ALLELIC SEQUENCE DIVERSITY AT THE HUMAN BETA-GLOBIN LOCUS

Stephanie Malia Fullerton

Somerville College
University of Oxford

Thesis submitted for the degree of Doctor of Philosophy
Michaelmas Term 1994

ALLELIC SEQUENCE DIVERSITY AT THE HUMAN ß-GLOBIN LOCUS

Stephanie Malia Fullerton
Somerville College, University of Oxford

Thesis submitted for the degree of Doctor of Philosophy, Michaelmas Term 1994

ABSTRACT

The investigation of genetic variation at the DNA sequence level in *Drosophila* has demonstrated the success of the approach for elucidating the impact of molecular, selective, and demographic evolutionary processes. However, sequence level investigation of human genetic variation has thus far been limited to a single locus, the mitochondrial DNA genome. In order to extend the analysis of DNA sequence variation to human nuclear loci, and thereby bring the power of sequence analysis to the investigation of human population origins and affinities, allelic sequence diversity at the human ß-globin gene was determined.

A 3 kilobase (kb) genomic region, encompassing the human ß-globin gene and associated flanking DNA, was examined. DNA sequence variation in this region was identified by direct sequence analysis of allele-specific PCR amplification products. In total, 173 chromosomes from five human populations from Melanesia, Indonesia, and Africa, were investigated. Thirty-four polymorphisms were identified: 26 silent nucleotide differences, 2 non-silent (ß haemoglobinopathies) variants, and 6 small insertion/deletion polymorphisms. These 34 polymorphisms were inherited as 53 different alleles, each defined by a unique sequence haplotype. Only 10 alleles were identified when the same genomic region was investigated by restriction enzyme analysis.

The allelic sequence diversity identified suggested much about the influence of molecular, selective, and demographic phenomena. The unexpectedly high level of sequence haplotype polymorphism was found to be consistent with the pronounced impact of interallelic recombination in sequences 5' of the ß-globin gene. Despite the presence of selected ß haemoglobinopathies in some of the populations investigated, neither the genomic distribution nor the frequency of silent polymorphic sites was found to be inconsistent with that expected of a neutrally-evolving locus. Geographic differences in ß-globin diversity were therefore attributed to the effects of mutation in partially subdivided populations as well as in specific cases reduction in population size and gene flow, rather than the influence of natural selection, and evaluated with respect to currently debated hypotheses concerning human evolutionary origins.

The investigation of DNA sequence variation at the human ß-globin locus has therefore contributed new insights into the evolution of the locus in human populations and provided an important new class of data with which to test hypotheses concerning the genetic origin of *Homo sapiens sapiens*.

"I look forward with no little anxiety to the time when Henslow, putting on a grave face, shall decide on the merits of my notes. If he shakes his head in a disapproving manner: I shall then know that I had better at once give up science, for science will have given me up."

– Charles Darwin, 1836 (Desmond & Moore, 1991, p.183)

ACKNOWLEDGEMENTS

I have, on occasion, marvelled at the number of individuals and organisations cited on the typical DPhil acknowledgements page and wondered how on earth the author could have so many people to thank. At the end of the writing of my own dissertation, however, I now appreciate the extent to which a successful programme of research relies on the expertise, support, and guidance of numerous friends and mentors. My list, too, is long.

For their generosity with precious DNA samples I'd like to acknowledge Mary Ganczakowski, Doug Higgs, and Adrian Hill, of the Institute of Molecular Medicine, as well as Prof ASM Sofro, of the Inter University Center for Biotechnology, Gadjah Mada University, Yogyakarta, Indonesia. Thanks are also owed to Richard Hudson, of the University of California at Irvine, who readily provided our laboratory with copies of extremely useful computer programmes for the analysis of DNA sequence data, as well as to Rosalind Harding, who figured out how to use them. Jonathan Flint helped analysis as well, by providing me with a computer account in which to work. Other (everyday) computing needs were met, time and time again, by Jeremy Martinson, Liz Cowe, and Tony Brett.

Experimentally, my research benefited tremendously from the (largely informal) assistance of a large number of colleagues at the Institute of Molecular Medicine in Oxford, many of whom sacrificed hours of their time to advise and coach me in the nuances of PCR, cloning and sequencing (as well as the very occassional Southern Blot!). I am grateful to Rhona Borts, Jonathan Flint, Mike Francis, Erika Hagelberg, Georgina Hall, Janette Lamb, Jeremy Martinson, David Neil, Ann Oakenfull, John Old, Robin Olds, Katherine Robson, Ed Smith, and Sweelay Thein in particular for their expert guidance.

A very great measure of appreciation goes to Jacquie Bond, who contributed to the sequencing studies reported in this thesis, and who continues to do experimental work on the ß-globin locus (with Julie Schneider and others). This study of ß-globin sequence diversity would have been far poorer had she not arrived on the scene.

Major emotional support was provided by friends both inside and outside the lab. Thanks go to Ann, Janette, Jemma, Jacquie, David, Peter, Emma, Susanne, Jen, and most especially, Tessa. I am indebted to Tessa, as well, for her careful reading of the countless draft chapters I generated. A big 'mahalo' to Dad and Amy in Hawaii, who put up with my computing throughout the month of September, and Mom and Chris, who didn't have to, but were also very supportive. Special thanks also to Dad for his own PhD stories, which kept me going during the darkest days of the write-up.

Many thanks go to my biological anthropology supervisor, Tony Boyce, for introducing me to the field of human population genetics in the first place (via the taught MSc in Human Biology) and then cunningly suggesting that I consider doing research at the Institute of Molecular Medicine. It is a decision I have never once regretted. Tony also provided key advice and support throughout the period of research, particularly in the final stages; I would not have been so easily able to prepare my thesis for submission without the material resources he placed at my disposal.

Finally, I could not have completed this thesis without the constant interest, support, and criticism provided by John Clegg, my laboratory supervisor, and Rosalind Harding, our research group's senior postdoc. John encouraged me in my laboratory work and made the necessary equipment and facilities available to me. He also secured additional funding support when my original grant from the Rhodes Trust ran out. Rosalind introduced me to many of the analyses I incorporated into the thesis and critiqued my own (often primitive) attempts at population genetics. I owe them both a great deal. Unfortunately all I can really do is say thanks.

TABLE OF CONTENTS

ALLELIC SEQUENCE DIVERSITY

AT THE

HUMAN ß-GLOBIN LOCUS

Chapter 1

Investigation of Genetic Variation at the Human ß-Globin Locus: Study Background and Design

Population geneticists have studied the molecular basis of naturally-occurring genetic variation for nearly 30 years, with the aim of understanding the processes responsible for evolutionary change. The earliest surveys of electrophoretic protein variation, in the fruit fly species *Drosophila pseudoobscura* (Hubby & Lewontin, 1966; Lewontin & Hubby, 1966), demonstrated the presence of extensive allelic polymorphism at numerous genetic loci. As these and subsequent analyses showed (Lewontin, 1985), protein-coding genes vary to differing degrees, and the level and nature of polymorphism detected at any particular genetic locus often differs considerably from population to population, as well as from species to species. The existence of pervasive genetic diversity, in organisms ranging from *Drosophila* to humans, was quite unlike anything previously anticipated and demanded suitable explanation in terms of molecular, selective, and demographic factors.

Two of the most important evolutionary forces with the potential for affecting variation in populations are natural selection and random genetic drift. Prior to the advent of electrophoretic study, it was assumed that most variation was adaptive and therefore maintained in populations by natural selection. The impact of genetic drift (i.e. changes in allele frequency from generation to generation resulting from the random sampling of gametes during zygote formation) was considered minimal. However, with the discovery of so much variation at the protein level, whose adaptive significance was difficult or impossible to infer, theoreticians reconsidered the role that genetic drift might play in the maintenance of observed variability. In 1968, Kimura, drawing on the work of Wright (who, over 30 years earlier, had emphasised the importance of random genetic drift for evolutionary change), redefined population genetic enquiry with the introduction of the neutral mutation hypothesis.

Genetic Variation and the Neutral Theory of Molecular Evolution

The neutral theory of molecular evolution (Kimura, 1968; 1983) suggests simply that most evolutionary changes (i.e. the fixation of mutant substitutions at the molecular level) come about as a consequence of random genetic drift. This straightforward prediction is based on assumptions regarding the phenotypic importance of particular nucleotide changes. Kimura argued that most mutations are either strongly deleterious, therefore apt to destroy or significantly compromise gene function, or selectively neutral, with little or no effect on genetic fitness. As strongly deleterious changes are eliminated rapidly by purifying selection, their occurrence has little impact on the nature of genetic variation in a population. Selectively neutral mutations, however, persist and it is these differences, which arise at random and drift by chance to fixation or loss, which determine evolutionary change.

In this context, intraspecific DNA polymorphism, a transient phase in long-term molecular evolution (Kimura, 1983), is determined by population size and mutation rate. The effects genetic drift, like any random sampling process, depend on the size of the sample. Therefore, when the number of breeding individuals in a population (i.e. the 'effective' population size) is small, the likelihood that the frequency of a particular allele* will change as a consequence of drift is large, whereas when the effective population size is large, drift has a proportionately smaller impact. The overall level of polymorphic variation in a population is determined by the rate of introduction of new alleles relative to the effects of drift. A high rate of mutation and/or recombination in a large population will result in large amounts of standing polymorphism because new alleles arise more quickly than old alleles are fixed or lost by drift. Once an equilibrium between the effects of mutation/recombination and drift is attained, the level of polymorphism in the population remains constant; the on-going turnover of alleles, hence variable sites, present in the population results in long-term molecular change.

---

* While the term 'allele' can be, and is often, used in this context to refer to one of the four bases present at a single polymorphic nucleotide position, here the term is understood to refer to a larger genomic region, encompassing part or all of a gene and generally containing within it one or more polymorphic nucleotide sites.

Testing the Neutral Mutation Hypothesis

The evolutionary basis of genetic variation found within and between populations can be assessed according to the predictions of the neutral mutation hypothesis. For example, the level of polymorphism observed in a single population (measured either as the number of alleles, or the number of polymorphic nucleotide positions) may vary considerably from locus to locus due to different rates of mutation and recombination. Similarly, the level of variation at a single genetic locus can differ from population to population as a consequence of differences in susceptibility to genetic drift (i.e. different effective population sizes). If the genes under investigation are evolving neutrally, however, *relative* differences in diversity among populations will be the same irrespective of the absolute level of polymorphism in any particular population (i.e. diversity in population A will be 50% higher than in population B at each locus investigated, no matter what amount of variation is observed in population B). If unexpected interlocus heterogeneity is observed, the influence of some other factor, such as natural selection, acting to increase or eliminate variation at a locus in one or more populations, is suggested.

Differences in the nature, rather than the level, of observed genetic variation among populations can also be assessed according to neutral theory expectations. When gene flow between populations is restricted, new genetic variants, arising either from mutation to, or recombination of, existing alleles, attain polymorphic frequencies while older alleles are lost via random genetic drift; as a consequence, significant differences in the genetic composition of the subdivided populations accrue. The extent to which populations no longer share similar types and frequencies of alleles is described as their genetic distance. If neutral, variation at all loci should indicate the same relative genetic distances among populations, differentiation which reflects both the time that drift (and mutation or recombination) has been acting independently in subdivided populations, as well as changes in population size which may have occurred since the populations diverged. Loci which reflect different population relationships, however, suggest the effects of natural selection, which has either influenced patterns of diversity

in some populations and not others or has rendered polymorphism in divergent populations more similar than they would have been by chance.

At the time the neutral mutation hypothesis was proposed, there were comparatively few polymorphism data available with which to assess the level or pattern of genetic variation in populations with respect to the predictions of the neutral theory, and the data which were available (surveys of electrophoretic protein variation) were not sufficient to discriminate the evolutionary processes of interest (Lewontin, 1985). The unsuitability of electrophoretic data derived principally from its low level of genetic resolution. Protein electrophoresis detects only a proportion of genetic differences present at the DNA level, those changes which result in amino acid substitutions. The possible influence of selection on such non-silent variation makes the relationship of the observed polymorphism to underlying genetic variability unclear, and renders any assessment of the level of polymorphism at different loci and/or in different populations effectively meaningless. The likely influence of selection on allelic variants also makes it impossible to distinguish whether patterns of genetic variation in populations, as revealed by allele frequency distributions, are concordant because of recent divergence from a common ancestral population or because of shared susceptibility to selective effects.

DNA Sequence Investigation of Genetic Variation

The introduction of recombinant DNA technology made possible the direct assessment of genetic variation at the DNA sequence level. The first such investigation, of the alcohol dehydrogenase locus (*Adh*) in *Drosophila melanogaster* (Kreitman, 1983), demonstrated the extent of genetic variation missed by electrophoretic analyses as well as the power of DNA sequence data in the investigation of evolutionary phenomena. In that study, 11 *Adh* alleles, associated with one of two known electrophoretic alloenzyme variants (fast, *Adh-f*, and slow, *Adh-s*), were cloned, sequenced, and compared. In addition to the single previously-identified nucleotide substitution at the locus (a mutation which results in a threonine to lysine amino acid replacement in the alcohol dehydrogenase enzyme), nearly 50 other length and point polymorphisms were

identified in the 2.6 kilobase- (kb) long genomic region investigated. The overall level of genetic diversity at the *Adh* locus, and the distribution of nucleotide polymorphism among different genomic regions and alleles, provided new insights into the factors responsible for the evolution of the gene in *D. melanogaster* populations.

The majority of the sequence polymorphisms identified at the *Adh* locus, for instance, were found in non-coding flanking and intervening (intron) genomic regions, in accord with the predictions of the neutral mutation hypothesis. The selectively deleterious nature of mutations which lead to amino acid replacements was further confirmed by the fact that only 1 of the 14 nucleotide variants observed in the exons of the *Adh* gene resulted in an amino acid change. The high level of remaining 'neutral' polymorphism at the locus was found to be consistent with an effective population size for the *D. melanogaster* species of several million (the 11 alleles were drawn from five different *melanogaster* populations from around the world). From the distribution of nucleotide polymorphism among the 11 alleles investigated, Kreitman was also able to infer that the slow allele was ancestral to the fast allele (the *Adh-f* alleles were, as a group, less polymorphic than the *Adh-s* alleles) and that at least two independent recombination events had occurred at the locus in the course of its evolution. These latter conclusions were supported by subsequent phylogenetic analysis of the same allelic sequences (Stephens & Nei, 1985).

Polymorphism in different regions of the *Adh* gene was also subsequently re-evaluated with respect to patterns of interspecific divergence between *D. melanogaster* and *D. simulans* (Kreitman & Aguadé, 1986b). As mentioned above, the neutral theory suggests that polymorphism is a transient phase in long-term sequence evolution (Kimura, 1983). As a consequence, genomic regions in which high levels of interspecific sequence divergence are found are also expected to display high levels of polymorphism within species. Kreitman, in collaboration with Aguadé and later Hudson (Hudson, Kreitman, & Aguadé, 1987), was able to demonstrate that significantly higher than expected levels of polymorphism were present at the *Adh* locus in *D. melanogaster* when interspecific sequence divergence was taken into account. The presence of this excess polymorphism was attributed to the effects of genetic 'hitchhiking' (Maynard

Smith & Haigh, 1974) to a selectively-maintained balanced polymorphism. DNA sequence analysis of allelic variation at the *Adh* gene in *D. melanogaster* therefore permitted the first direct assessment of neutral theory predictions and provided compelling evidence for the effects of natural selection on the locus, without recourse to extensive population sampling or interlocus genetic comparison.

Experimental and analytical approaches to the analysis of genetic variation pioneered by Kreitman and his colleagues have revolutionised population genetic investigation of *Drosophila melanogaster* and related species. Numerous *Drosophila* loci have now been examined at the DNA sequence level, often in two or more species, and the results suggest the influence of a wide variety of evolutionary processes (*Adh*, for instance, is the only fruit fly locus for which there is convincing evidence of the effects of balancing selection on polymorphism). These studies, and the data they have generated, have also motivated the development of other sequence-based statistical tests of the neutral mutation hypothesis (i.e. Tajima, 1989a; McDonald & Kreitman, 1991) as well as new methods of assessing the distribution of variation within and between populations using DNA sequence data (i.e. Strobeck, 1987; Lynch & Crease, 1990; Hudson, Boos, & Kaplan, 1992). As a result, today we are in a much better position to evaluate successfully the contributions of mutation, recombination, migration, random genetic drift, and natural selection to the level and distribution of genetic diversity in populations.

Human Population Genetic Investigation

The investigation of genetic variation in human populations has, to a large extent, lagged behind equivalent study in *Drosophila*. Human populations are less amenable to sampling than fruit flies and experimental manipulation is, in many cases, more complex. This is particularly true for the analysis of diploid nuclear DNA regions, where the use of isochromosomal species lines to simplify analysis of maternal and paternal alleles present at the same genetic locus (a strategy commonly employed in the analysis of *Drosophila* species), is not a viable option. In addition, the emphasis in

human molecular genetics on the identification of mutations responsible for inherited genetic disorders has meant that, in most instances, studies of DNA variation have been motivated by the search for markers suitable for mapping the chromosomal location of affected genes, rather than a desire to understand the evolutionary processes responsible for observed genetic differences.

Nevertheless, motivated largely by a desire to infer the nature and timing of human evolutionary origins,* a number of studies have attempted to identify and explain genetic variation present in human populations. Two approaches have dominated investigation: the combined assessment of protein or DNA polymorphism present at multiple unlinked loci sampled from across the entire human genome, and DNA-level analysis of allelic variation present at individual genes. Whole genome approaches are favoured by some investigators because they average genetic patterns over many independently-evolving loci and therefore maximise the likelihood of identifying similarities and differences which derive from population history, rather than from locus-specific influences. The need in such studies to examine large numbers of loci, however, limits both the number of populations which can be investigated and the level at which variation can be assessed. Individual genetic loci, in contrast, can be studied in greater detail, so that not only the frequencies but also the identities of alleles present in particular populations can be determined and used to assess genetic affinities. As discussed above, this detail can help the investigator gauge the extent to which variation at the locus is neutral and, hence, likely to reflect demographic rather than selective phenomena.

The Simultaneous Investigation of Genetic Variation at Many Unlinked Loci

Multi-locus analyses of human genetic variation have focused primarily on two kinds of genetic data: protein, i.e. 'classical' electrophoretic enzyme and blood group markers, and DNA, restriction fragment length polymorphisms (RFLPs) and microsatellite loci. In

---

* This motivation, in fact, represents a fundamental difference between most recent human and *Drosophila* studies: the latter are rarely presented in terms of population origins and affinities. One exception is the study of a Bogota sample of *Drosophila pseudoobscura,* in which the timing of population isolation is estimated (Schaeffer & Miller, 1991).

these studies, genetic distances between pairs of populations are calculated from observed allele frequency differences and then used to infer an overall pattern of genetic relationship among populations (most often presented as a phylogenetic tree). Despite substantial differences in the way in which genetic distances are calculated and the resulting genetic relationships described, conclusions from such investigations have been remarkably concordant. Two recent surveys of classical genetic markers (Cavalli-Sforza et al., 1988; Nei & Roychoudhury, 1993), for example, considered different loci and populations and employed very different analytical strategies, yet each concluded that the largest genetic differences are found between African and non-African populations. These studies did, however, disagree on the nature of genetic relationships among non-African populations. Whereas Cavalli-Sforza et al. (1988) found, in accord with major linguistic affiliations, that non-African populations subdivided into Northern Eurasian and Southeast Asian groups, Nei & Roychoudhury (1993) inferred a Caucasian/non-Caucasian secondary split which they claim is better supported on statistical grounds and more consistent with archaeological and morphological data.

As classical polymorphisms do not reflect reliably variation present at the molecular level, attempts have been made to augment electrophoretic studies with the direct investigation of DNA diversity. The largest multi-locus analysis of DNA polymorphism to date has involved the assessment of 100 RFLPs, from 42 nuclear DNA genes and 31 anonymous segments, in 5 populations from Africa, Asia, Europe, and Melanesia (Bowcock et al., 1987; 1991a; 1991b). In these studies, the presence or absence (+ or -) of restriction enzyme recognition sequences in particular locales was used to calculate genetic distances among populations. The results confirmed the genetic differentiation of African and non-African populations. A comparison of the allelic status of 79 RFLPs in humans and closely-related primate species has since suggested that the observed RFLP variation is most consistent with an African origin for human populations (Mountain & Cavalli-Sforza, 1994). Although the European sample was more polymorphic than either of the African populations investigated* (Bowcock et al.,

* Most models of population expansion predict that the greatest amount of polymorphic variation would be found in the 'ancestral' population (discussed in Chapter 6).

1991b), this finding has been explained as an artefact associated with the initial identification of the RFLPs in European populations (Mountain & Cavalli-Sforza, 1994).

Bi-allelic RFLPs afford only limited resolution of the allelic variation likely to be present at most genetic loci.* For this reason, variation in polymorphic microsatellite loci (CA repeats, in particular) has also been investigated. A preliminary examination of 30 microsatellite loci in 10 individuals from each of 14 different human populations from Africa, Europe, East Asia, Oceania, and America (Bowcock et al., 1994), demonstrated the clustering of individuals from similar geographic regions when pairwise interindividual differences (calculated as the proportion of alleles shared over all loci) were taken into account. Genetic distances among populations rather than individuals were also investigated, and in common with other whole genome analyses, a statistically-significant difference between African and non-African populations was observed. Unlike nuclear DNA RFLPs, however, the highest level of variation was observed in the African populations. Bowcock and colleagues suggest that the large number of alleles identified at microsatellite loci (approximately 7 alleles per locus were identified in most major continental regions) render ascertainment bias a less serious problem for studies of this kind.

DNA sequence analysis has not been applied to a concerted examination of population variation at multiple genetic loci. One recent investigation of published cDNA and genomic sequences, identified in the course of clinical investigation and made accessible by virtue of their deposition in the DNA sequence database GenBank (Bilofsky & Burks, 1988), has provided certain preliminary insights into the nature of human variation at the DNA sequence level (Li & Sadler, 1991). That study, which compared two or more sequences reported for 49 different nuclear DNA loci, found an average level of pairwise genetic difference between alleles of 0.08% at synonymous nucleotide sites. This rather low level of nucleotide diversity (an order of magnitude lower than similar estimates for the fruit fly species *Drosophila pseudoobscura*) is consistent with a relatively small long-term effective population size for the human

---

* While combined consideration of linked sites as 'RFLP haplotypes' can circumvent this drawback, RFLPs have only seldom been used in this way in multi-locus investigations.

species. As the data were not collected in a population context, however, differences in the level of diversity in particular populations could not be investigated.

Single Locus Investigations

Genetic variation at individual loci, although susceptible to the confounding influence of locus-specific processes, is more amenable to fine-scale allelic investigation using RFLP haplotype or DNA sequence analysis. The allelic discrimination afforded by such techniques provides increased resolution for standard (i.e. frequency-based) analytical approaches and allows population geneticists to take advantage of additional phylogenetic methods for the investigation of genetic diversity (reviewed in Hudson, 1990). The most extensive survey of nuclear DNA variation at a human genetic locus to date has been the analysis of RFLP haplotype polymorphism in the ß-globin gene cluster on chromosome 11 (Wainscoat et al., 1986). Five RFLP sites were assayed in each of 600 chromosomes and 14 (of the 32 theoretically possible) haplotypes identified in the 10 populations investigated. Three of the 14 haplotypes accounted for the majority of alleles observed in non-African populations. The most common haplotype in the 2 African samples, however, was not found in any of the non-African populations investigated. Wainscoat and colleagues argued that these patterns were consistent with the migration of a founder population out of Africa which subsequently gave rise to all non-African populations.

Rare alleles contribute very little to frequency-based investigation of genetic variation in populations. To make fuller use of the available data, ß-globin RFLP haplotype diversity was re-interpreted subsequently, in phylogenetic terms (Long et al., 1990; Chen et al., 1990). Phylogenetic analysis involved two stages. First, a network of presumed evolutionary relationship among observed haplotypes was constructed according to a principle of maximum parsimony (reviewed in Stewart, 1993). Once the allelic genealogy was identified, relationships among populations were derived from the shared presence of derived haplotypes or shared absence of ancestral haplotypes. The tree of population relationship inferred by this kind of analysis provided a tentative indication of the sequence of mutational and recombinational events that occurred at

the locus during the course of its evolution and confirmed that the observed variation was consistent with a human African origin. Difficulties in distinguishing the effects of mutation and recombination made the process of inferring phylogenetic relationships among the haplotypes exceedingly complex, however. Indeed Long et al. and Chen et al. arrived at different conclusions about the nature of population relationships because of different interpretations of RFLP haplotype genealogy.

Similar investigation of RFLP haplotype diversity at the human α-globin cluster on chromosome 16 has been undertaken (Higgs et al., 1986, Martinson et al., in preparation) and the relationship of the observed variation to human population affinities and evolutionary history is currently under investigation (A. J. Boyce, personal communication). The distribution of normal and thalassaemic α- and ß-globin RFLP haplotypes in Melanesia, Micronesia, and Polynesia, have provided important insights into the genetic affinities of populations in these geographic regions (Hill, O'Shaughnessy, & Clegg, 1989), as well as the impact of malarial selection on the distribution of haemoglobinopathies throughout Oceania and other equatorial areas (Flint et al., 1993a; 1993b). The only other published population genetic study of allelic variation at a nuclear locus has been the analysis of sequence diversity in a 300 basepair (bp) segment of the human pseudoautosomal boundary region (Ellis et al., 1990). That investigation, of 57 Y and 60 X chromosomes in 10 populations, identified 5 X-chromosome nucleotide polymorphisms; only 1 of these 5 polymorphisms was observed in certain (African) Y chromosomes. Two major X chromosome haplotypes predominated in most populations, while a third recombinant haplotype was observed in African samples only. No frequency-based or phylogenetic analysis of these haplotypes was reported.

For the last 10 years, however, human population genetic analysis has been dominated by the investigation of variation at a different genetic locus, the human mitochondrial DNA (mtDNA) genome. Mitochondrial DNA, unlike nuclear DNA, is haploid (hence, non-recombining) and strictly maternally-inherited (Giles et al., 1980); these features make mtDNA variation easier to analyse and interpret phylogenetically. In addition, because the overall rate of mutation of mtDNA is almost an order of magnitude higher

than equivalent nuclear loci (Cann, Stoneking, & Wilson, 1987) there is much more, potentially informative, mtDNA variation within and between human populations available for analysis. The first investigation of mtDNA variation in human populations (Johnson et al., 1983) involved the analysis of 40 mtDNA RFLP sites in 200 samples from Africa, Asia, Europe, and the New World. Thirty-five mtDNA haplotypes were identified, haplotype diversity was greatest in African populations, and significant differences between African and non-African populations were suggested. Johnson and colleagues explained differences in the level of mtDNA diversity in terms of an increased rate of mutation in African populations and concluded that the observed variation was consistent with an Asian origin of anatomically modern humans (*Homo sapiens sapiens*).

That investigation was soon superseded by high resolution RFLP analysis of 147 mtDNA genomes sampled from African, Asian, European, Australian, and Melanesian populations (Cann, Stoneking, & Wilson, 1987). Cann and colleagues identified 133 distinct mtDNA haplotypes using restriction enzymes which cut the mtDNA genome at 467 independent positions; only 7 of these 133 types were found in more than 1 individual. As the large number of unique mtDNA alleles identified made consideration of population affinities in terms of haplotype frequencies impossible, a phylogenetic tree connecting the mtDNA types was constructed. The resulting gene genealogy suggested a common African origin for all mtDNA alleles (one of the two primary branches of the tree led exclusively to African mtDNA types while another led to both African and non-African alleles). Average pairwise differences between sequence haplotypes were also greatest among African alleles, consistent with an African origin for the observed variation. By measuring sequence differences between haplotypes in either half of the genealogy, and comparing them to the estimated rate of sequence divergence at the locus, Cann and colleagues estimated that the African mtDNA ancestor lived 140,000 to 290,000 years ago.

These mtDNA findings excited great interest when first presented, largely as a consequence of the unexpected recency of human population divergence implied by the observed allelic differences. Existing palaeontological evidence was felt by many

(i.e. Wolpoff, 1989) to be consistent with a far earlier African origin, of the pre-human species *Homo erectus*, approximately 1 million years ago (geographically-dispersed *erectus* populations evolved subsequently, *in situ*, into *H. s. sapiens*). Other criticisms focused on the study's reliance on RFLP rather than DNA sequence data (Wolpoff, 1989), the use of African-American samples, as opposed to mtDNA from populations living in Africa (Excoffier & Langaney, 1989), as well as the method of rooting the tree and calibrating the rate of sequence divergence at the locus (Saitou & Omoto, 1987). Most of these criticisms were addressed by the subsequent investigation of DNA sequence variation in 2 regions of hypervariable mtDNA from 189 individuals, including 121 native Africans (Vigilant et al., 1991). Using new phylogenetic methods, that study confirmed the African origin of mtDNA diversity and estimated the age of the nearest common ancestor at between 166,000 and 249,000 years, providing "the strongest support yet" for a recent African origin of modern humans populations.

Shortly after the publication of the Vigilant et al. (1991) paper, however, systematists David Maddison and Alan Templeton, in separate articles (Maddison, 1991; Templeton, 1992) questioned the statistical reliability of the phylogenetic tree generated using the mtDNA sequence data. Both were able to demonstrate the existence of more parsimonious trees, often with non-African roots, suggesting that the origin of mtDNA sequences could not be placed with certainty in Africa. As a consequence, Stoneking (author on both the Cann, Stoneking, & Wilson, 1987 and the Vigilant et al., 1991 papers) conceded that the "available sequence data are insufficient to statistically resolve the geographic origin of human mitochondrial DNA" (Hedges et al., 1992). Although Cann maintains that the existence of significantly greater genetic variation in African populations is further evidence of the likely African origin of mtDNA lineages (Gibbons, 1992), both this conclusion and the estimated times of sequence divergence have been called into question by Templeton (1993). It is likely, therefore, that the debate concerning interpretation of mtDNA variation will remain unresolved until equivalent data from other human (i.e. nuclear) loci are obtained and analysed.

Allelic Sequence Diversity at a Human Nuclear Locus

There is little doubt that human population genetic enquiry would benefit enormously from the investigation of DNA sequence variation at nuclear DNA loci: whole genome approaches are currently too imprecise to permit meaningful conclusions to be drawn regarding the causes of genetic similarities and differences among human populations, and variation identified at the mtDNA locus has, for a variety of reasons, eluded easy interpretation. Technical difficulties associated with the isolation and analysis of nuclear DNA alleles, however, have meant that no analysis of allelic sequence variation at a human nuclear locus has been attempted. With the advent of new methodology and the recognition that nuclear DNA sequence data are a fundamental prerequisite for the genetic investigation of human evolutionary origins (Hedges et al., 1992), the detailed examination of a human nuclear locus at the sequence level has become both feasible and imperative.

The aims of the research presented in this thesis were, therefore, two-fold: first, to initiate sequence-level investigation of a human nuclear locus in a population context, thereby establishing the feasibility of such study at potentially any human genetic locus, and second, to evaluate the variation identified in terms of likely molecular, selective, and demographic effects. Of interest in the latter regard was the nature and level of nucleotide and sequence haplotype diversity present at a nuclear gene, its distribution within and between populations, and the relationship of observed genetic patterns to results reported for other loci, particularly mtDNA. Although it was recognised that analysis of a single nuclear locus was as unlikely to 'answer' the question of human origins as mtDNA data, implications of the observed variation for human evolutionary origins were also examined.

Relevant Considerations in the Design of the Study

As no population genetic investigation of a human nuclear locus at the DNA sequence level had been undertaken at the start of research, there were few precedents available with which to guide gene choice and study design. An immediate consideration was

which locus to investigate, and what sort of genomic regions (i.e. coding or non-coding, single copy or repetitive) to include in the segment to be analysed. Analysis of a genetic locus that was, to some extent, already well-characterised was clearly advantageous, as the relationship of polymorphism to gene structure and function is relevant to the investigation of many neutral theory predictions. Previous analysis of the locus in a population context was also desirable, for it would allow comparison of results obtained using different detection strategies, as well as serve to extend results beyond the confines of the immediate investigation.

As noted above, the most extensively investigated nuclear locus in human population genetic analysis has been the ß-globin gene cluster. This 'locus', as assayed by RFLP haplotype analysis, is approximately 35 kb in length - too large to be studied by high resolution methods like DNA sequence analysis. The gene cluster is, however, comprised of five smaller functional loci (and a closely related, but non-expressed pseudogene) which code for the ß chains of the human haemoglobin protein during different stages of development (Karlsson & Nienhuis, 1985). The most extensively studied of these loci is the human ß-globin gene, which produces the adult-expressed haemoglobin ß protein chain. This gene is affected by a large number of mutations which alter the structure and/or synthesis of the ß protein chains and lead to severe haemoglobinopathies such as sickle cell anaemia and ß thalassaemia (Antonarakis, Kazazian Jr., & Orkin, 1985). As a consequence of intense clinical scrutiny of the molecular basis of these disorders, much is already known about sequence variation at the locus, although not in any great detail, and not in a population context. These features, however, combined with the known influence of natural selection on the locus (Flint et al., 1993a; 1993b), make it an appropriate subject for in-depth, population-based analysis.

Another important consideration was the size of the genomic region to be investigated. DNA sequence analysis is labour-intensive and largely infeasible for the investigation of genomic regions larger than 5 kb in length (at least when manual sequencing is employed, as was the case here). While the maximum size limit of the region to be investigated was indicated by methodological constraints, the minimum size limit was

less certain. Sequence investigation of the mtDNA locus found nearly 200 variable nucleotide sites in a genomic region less than 1 kb in length (Vigilant et al., 1991). This 'hypervariable' segment of the mtDNA genome is, however, likely to be evolving more than 50 times as quickly as the ß-globin gene region (Savatier et al., 1985) so that many fewer polymorphisms would be expected to be observed in a similarly-sized segment of ß-globin DNA. Indirect measures of ß-globin diversity, based on RFLP analysis of genetic variation in non-coding DNA (Jeffreys, 1979; Kazazian Jr. et al., 1983), have estimated that anywhere from 1 in 100 to 1 in 1000 nucleotide sites may vary polymorphically at the locus. Together, these findings suggested that several kilobases, preferably of non-coding DNA,* would need to be analysed to ensure that some nucleotide polymorphism would be detected.

The ß-globin locus is, in its entirety, approximately 1.5 kb in length. It is comprised of 3 exons and 2 introns, and the introns make up the majority of the gene (Lawn et al., 1980). The size and overall composition of the locus therefore make it ideal for DNA sequence-level investigation. A nearly equivalent amount of non-coding flanking DNA (800 bp 5' and 600 bp 3' of the gene) was also included in the segment to be analysed. The inclusion of flanking DNA was motivated by a number of considerations. A comparison of levels of interspecific sequence divergence in different genomic regions has suggested that introns may in fact be constrained functionally in relation to non-coding flanking DNA and pseudogenes (Li, Luo, & Wu, 1985). DNA sequence analysis of flanking DNA therefore permitted comparison of levels of polymorphism in the relevant genomic regions. In addition, although previous investigation of the ß-globin locus has included DNA 5' of the gene (Moschonas, de Boer, & Flavell, 1982; Semenza et al., 1984a; 1984b; Cheboune et al., 1988) no sequence-level analysis has been reported for the 3' flanking region. By including approximately equal amounts of 5' and 3' DNA, the nature and level of polymorphism on either side of the ß-globin gene could be also examined.

A further consideration in the design of the study was whether to investigate DNA

---

* The neutral theory predicts that the most variation will be found in the least functionally constrained regions.

sequence variation in 'normal' genes only (i.e. genes known not to be affected by a mutation or deletion which might result in a haemoglobinopathy) or to include in the analysis genes affected by such mutations. There is ample evidence to suggest that the geographic distribution and frequency of such ß haemoglobinopathies is strongly determined by the protection affected alleles afford heterozygous carriers against malarial infection (reviewed in Flint et al, 1993a; 1993b). For this reason, population genetic investigation of the larger ß-globin gene cluster has focused exclusively on non-affected chromosomes, so that as 'neutral' a pattern of polymorphism as possible could be determined. Leaving aside the question of whether in fact this assumption is valid, such treatment is in many ways inaccurate because the allelic composition of the population investigated is incompletely represented. Thus, it was decided that variation in both normal and affected genes should be examined. The complete monomorphism of affected alleles (suggested by earlier analyses [e.g. Trabuchet et al., 1991] and confirmed by the sequence investigation conducted here) also emerged an important analytical tool: by assuming the sequence haplotype of the affected allele in heterozygote carriers, the character of associated normal alleles could often be inferred unambiguously.

One of the most important aspects of the study was the investigation of allelic sequence variation in a *population* context. While some information about levels and patterns of polymorphism present at the human ß-globin locus could have been inferred from variation observed in a single population, it is unlikely that the influence of molecular, selective, and demographic processes would have been identified without similar investigation of diversity in other populations. It is possible, for instance, to distinguish genetic patterns which arise as a consequence of molecular phenomena (i.e. hotspots of mutation and/or recombination) from those which result from population specific processes (i.e. selection or demographic phenomena) by comparing the nature of genetic variation identified in several different populations. In order to distinguish selective and demographic influences, variation in populations in which the ß-globin locus is likely to be subject to malarial selection (i.e. those populations known to possess endemic levels of malaria as well as high frequencies of ß haemoglobinopathies) were compared to populations free of apparent selection.

The choice of populations to study within these categories was governed to a large extent by time constraints and the availability of samples (researchers at the Institute of Molecular Medicine in Oxford have had a long-standing interest in the genetics of Melanesian populations, for instance, making that locale a logical starting point for regional investigation), as well as by the desire to explore variation at the locus in terms of human evolutionary origins. In order to address the latter question, populations in African and non-African locales were surveyed. In total five populations, three from Melanesia (two from the island archipelago Vanuatu, Espiritu Santo and Maewo, as well as an Eastern Highlands Province population from Papua New Guinea [PNG]), one from Indonesia (Palembang, Sumatra), and one from Western Africa (the Gambia), were investigated. This allowed for close examination of genetic similarities and differences within a fairly confined region of Island Southeast Asia (understood here to include Island Melanesia and PNG), and comparison of these patterns with intercontinental differences in genetic diversity.

The number of alleles to investigate in each of these populations was another important research consideration. Sample size can have a bearing on interpretations of the nature and extent of allelic variation in particular populations: if the sample size is small relative to the total number of alleles present at the locus, the variation identified may inaccurately represent the 'true' pattern of polymorphism in the population. As with the size of the genomic region to be investigated, however, lack of detailed information about the nature of DNA sequence level variation at the ß-globin locus made decisions regarding sample size adequacy difficult. Restriction enzyme analysis of the ß-globin gene cluster had previously identified four major RFLP haplotypes (Kazazian Jr. et al., 1983; Wainscoat et al., 1986), while DNA sequence analysis of thalassaemic ß globin genes also suggested the presence of four common ß-globin alleles (Orkin & Kazazian Jr., 1984; Antonarakis, Kazazian Jr., & Orkin, 1985). Small-sized population samples (average, 24 chromosomes per locale) appeared appropriate in light of these observations. Ideally, estimates of genetic variation in different populations would have been based on the examination of similar numbers of alleles in each locale; in practice, this proved not to be possible. Most of the analytical methods employed to

compare population variation took sample size differences into account.

A final, very important, consideration was how to identify the sequence haplotype of each allele in heterozygous individuals. The polymerase chain reaction or PCR (Saiki et al., 1985; Mullis et al., 1986), which greatly simplifies the isolation and analysis of defined genomic regions and is therefore ideal for population-based investigation, simultaneously amplifies DNA from both alleles present in a given individual's genomic DNA. Thus, if the individual is heterozygous at more than one site in the amplified region, the resulting DNA sequence cannot be interpreted in allelic terms and important information about variation at the locus is lost. The prospect of cloning individual alleles found in heterozygote individuals (i.e. the majority of those sampled) was impracticable, and for this reason, a number of alternatives to this basic strategy were investigated (discussed at length in Chapter 2). Indirect (i.e. non-sequencing) methods, which might provide the same level of detail about allelic variation at the ß-globin locus, but require fewer laboratory manipulations, were explored simultaneously.

The Questions Investigated

The allelic sequence diversity identified at the ß-globin locus and its significance, in terms of molecular, selective, and demographic influences, is discussed in the remainder of the thesis. Chapters 3 through 6 are organised to address a number of questions which, in line with the aims of research, are focused on elucidating the nature and level of nucleotide and sequence haplotype diversity present at the ß-globin gene, understanding the distribution of allelic polymorphism within and between the surveyed populations, and determining the relationship of observed genetic variation to previously reported findings. These questions, in order of chapter presentation, are:

- What proportion of nucleotide sites vary at the locus? How many different alleles (or sequence haplotypes) are present at the locus and at what frequencies are they found in particular populations? What is the relationship between nucleotide and sequence haplotype polymorphism, i.e. how are nucleotide polymorphisms distributed among alleles?

- Is the sequence variation identified consistent with the expectations of the neutral mutation hypothesis or has natural selection influenced the nature or distribution of nucleotide and/or sequence haplotype polymorphism in particular populations?

- What is the nature of genetic differences within and between populations? Do population differences in the level and/or pattern of sequence diversity result from the effects of mutation, migration, selection, or drift?

- And finally, in light of other answers, what do geographic differences in polymorphism imply about human origins and how do findings compare to those reported by other human genetic studies?

Chapter 2

Development of the Experimental Approach

As discussed in Chapter 1, the principal aim of research was to determine fully the nature of allelic diversity present at the human ß-globin locus. The best means of investigating variation at the nucleotide level is DNA sequencing. DNA sequence analysis can detect all polymorphic differences present in a particular segment of DNA, locate precisely where those differences occur, and identify the nature of each substitution or insertion/deletion event. The allelic linkage relationship of polymorphisms found at multiple variable sites in a given genomic region (the sequence haplotype) is also revealed by DNA sequencing.

DNA molecules can be sequenced in one of two ways: by base-specific chemical degradation of end-labelled single-stranded DNA (ssDNA) (Maxam & Gilbert, 1977) or by enzymatic incorporation of chain-terminating dideoxynucleotides during DNA synthesis (Sanger, Nicklen, & Coulson, 1977). Both methods involve multiple analytical steps and are time-consuming, although the recent introduction of commercial molecular biology kits for dideoxy sequencing has made the application of this technique routine. DNA sequencing has also been simplified by the use of *in vitro* amplification, via the polymerase chain reaction or PCR (Saiki et al., 1985; Mullis et al., 1986), for the generation of template DNA. Prior to the advent of PCR, obtaining the sequence of a particular genomic region, from even one individual, involved days or weeks of screening genomic libraries, cloning, and subcloning the fragment of interest (Sambrook, Fritsch, & Maniatis, 1989).

PCR, however, because it results in the simultaneous amplification of both alleles of diploid loci, presents new problems for population geneticists interested in studying nuclear DNA sequence variation. When multiple heterozygous sites are observed in the same sequence, the linkage phase of sites is impossible to distinguish and, as a

consequence, important information about haplotype diversity is lost. The use of isochromosomal lineages to render sequenced regions effectively homozygous prior to amplification, a common strategy in studies of *Drosophila* sequence variation, is not applicable to human investigation. Therefore, a reliable alternative experimental approach was required. Two methods of preparing allele-specific template for sequence analysis were investigated, and indirect methods of assessing DNA variation were also explored.

Sequencing Cloned PCR Products

Cloning provides a straightforward way of isolating allele-specific DNA for sequence analysis. This is because each clone derives from a single recombinant molecule and, therefore, reflects the nucleotide composition of one allele only. Cloning vectors also serve as important tools in both the isolation of ssDNA template and subsequent DNA sequence analysis. Filamentous phage M13 vectors (Messing, 1983), for instance, which can be isolated as strand-specific ssDNA, produce template suitable for immediate DNA sequence analysis. Similarly, use of the universal sequencing primer, an oligonucleotide complementary to vector DNA flanking the site of insertion, standardises sequence analysis of otherwise unique genomic regions.

PCR products are often difficult to clone (Shuldiner, Scott, & Roth, 1990; Crowe et al., 1991) however, and cloning even a single sample can be time-consuming. Moreover, if M13 is used, each sample must be cloned twice, into vectors which carry the cloning sites in opposite orientations (e.g. M13 mp18 and mp19) so that both strands of the cloned DNA are available for sequencing (analysis of the complementary strand is often required to confirm apparent sequence differences). Clearly, even if *Taq*-induced sequence artefacts are negligible (as has been suggested by some authors, e.g. Kwiatowski et al., 1991), so that sequencing multiple clones of the same PCR product is not required, cloning and analysis of 24 alleles (the average size of each of the population samples investigated here) can be prohibitively laborious.

Given the laborious nature of the cloning process, it was reasoned that simultaneous ligation and transfection of individually-amplified PCR products would greatly ease the burden of preparing large numbers of ssDNA templates for sequence analysis. In order to determine which isolated clone derived from which sample (a prerequisite for estimating allele frequencies in the population surveyed), a PCR-based sample labelling strategy was developed. The application of this strategy to the analysis of 36 chromosomes from Espiritu Santo, Vanuatu, is described here (further methodological details can be found in Appendix B).

Simultaneous Cloning of Uniquely-Labelled PCR Products

To unambiguously label amplification products for subsequent identification, unique combinations of modified oligonucleotide primers were used for sample-specific PCR amplification. Once cloned, primer modifications were identified by sequence analysis of the insert DNA.

PCR amplification of a 2.6 kilobase (kb) region including the ß-globin gene and associated flanking regions, was achieved using oligonucleotide primers homologous to the GenBank (Bilofsky & Burks, 1988) ß-globin consensus sequence (Appendix B). Primers, and amplification conditions, were then modified to allow for the incorporation of degenerate restriction enzyme recognition sites at the 5' ends of the oligonucleotides. Just inside these enzyme sites were placed additional, degenerate, bases: one (G, A, T, or C) in each of four 5' amplification primers, and two (GG, AA, TT, CC, GA, GT, GC, AG, AT, AC, TG, TA, TC, CG, CA or CT) in each of sixteen 3' amplification primers. These 'tagged' primers were then used, in all possible combinations (i.e. 4 x 16), to uniquely amplify 64 different genomic DNA samples.

Equal aliquots of the 64 amplification products were pooled, digested, and purified by gel electrophoresis onto DEAE-cellulose membrane (Scheicher & Schuell). DNA purification was absolutely essential for successful ligation, and DEAE purification proved the most reliable means of obtaining highly purified PCR product after restriction enzyme digestion.

Following purification, the pooled insert DNA was ligated with M13 mp18 and mp19 vectors and transfected into competent *E. coli*. Transfection efficiency averaged 5 x 10^3 plaques per μg vector DNA. Single-stranded DNA was prepared from 96 colourless (i.e. recombinant) mp18 (5'->3' strand) colonies and 96 mp19 (3'->5' strand) recombinants. Upon gel electrophoresis, 94 mp18 and 94 mp19 ssDNAs were found to contain a 2.6 kb insert. Both ends of each insert were sequenced to determine sample origin: in this way, 187 (99.5%) clones were unambiguously identified (Table 2.1). 58 (91%) of the 64 PCR products which were simultaneously cloned were isolated as ssDNA: 46 (72%) in mp18 and 48 (75%) in mp19. On average, two copies of each sample were identified (in one or both orientations), although the actual number varied from 0 to 10 in mp18, and 0 to 6 in mp19. Thirty-six of the 64 amplified samples were successfully cloned in both mp18 and mp19, and these were selected for investigation by DNA sequence analysis.

Sequence Variation Observed in Cloned ß-Globin Alleles

The first 600 basepairs (bp) of the 5'->3' strand of the 2.6 kb fragment were sequenced (Appendix B) from each of 36 mp18 clones (the ssDNA preparations chosen for analysis are underlined in Table 2.1). These data, which encompass nucleotides -800 to -200 bp 5' of the human ß-globin gene, are summarised in Figure 2.1.

Thirty point differences, two 1-bp deletions, and one 2-bp insertion, distributed more or less uniformly across the 600 bp region, were identified among the 36 cloned sequences. Twenty-six of the 33 variable positions identified (24 point differences and 2 deletions) occurred only once in the 36 samples surveyed. The remaining 7 differences (positions 145, 297, 319, 320, 379, 470, 508) were found at frequencies ranging from 6 to 83%. Twenty-seven (90%) of the 30 nucleotide polymorphisms observed were transitions, of which 16 (59%) were T->C changes, 7 (26%) were A->G differences, and 4 (15%) were G->A changes. Two of the 3 transversions observed were T->A replacements, while the other was a T->G difference. Both deletions involved the first A of an AA doublet, and the 2-bp (AT) insertion occurred at the 3' end

PCR Sample	Primer 5'	Primer 3'	mp18 ssDNAs	mp19 ssDNAs
1	A	AA	93	7
2	T	AA		73, 74
3	C	AA		79
4	G	AA	70	
5	A	AT	52, 60, 71, 80	20, 27
6	T	AT	17	17, 67
7	C	AT	68	10, 42, 45, 63
8	G	AT		13, 23
9	A	AC	61, 83	56, 59
10	T	AC	12	87
11	C	AC	15, 51	30, 37
12	G	AC		
13	A	AG		40, 55
14	T	AG	90	65
15	C	AG		
16	G	AG		78
17	A	TA	66, 75	75
18	T	TA	1, 9	
19	C	TA		1, 95
20	G	TA		6
21	A	TT	62	51
22	T	TT	25, 28, 45, 65	5, 91
23	C	TT	3, 18, 36, 63	26, 28, 32
24	G	TT	6	31
25	A	TC		48
26	T	TC	53	38
27	C	TC		86
28	G	TC	89	52
29	A	TG	16	80
30	T	TG	21	14, 35
31	C	TG		
32	G	TG	39	82
33	A	CA	24, 86	16, 92
34	T	CA	26, 32, 69	41, 72, 77
35	C	CA	22, 35, 41, 77	22, 44
36	G	CA	54, 64, 78	21, 33, 69
37	A	CT	11, 31, 33, 44, 48, 57, 59, 72, 84, 9	25, 36, 60, 81
38	T	CT	58, 96	2
39	C	CT	7	
40	G	CT	55	
41	A	CC	8	
42	T	CC	4	53
43	C	CC		
44	G	CC	37, 38	89, 96
45	A	CG		
46	T	CG		11, 19
47	C	CG	34, 43, 85	4, 24, 66, 93
48	G	CG		70
49	A	GA	95	
50	T	GA	76	18, 46, 49, 50, 64
51	C	GA	56	12, 83
52	G	GA	87	
53	A	GT		85
54	T	GT	13, 92	15
55	C	GT	19	68
56	G	GT	30	
57	A	GC	27, 79	39, 84
58	T	GC	23, 50	3, 47, 61, 76
59	C	GC		
60	G	GC	20	
61	A	GG	14, 29, 49, 73, 94	29, 34, 57, 58, 90, 94
62	T	GG	10, 40, 47, 81, 82	8, 54
63	C	GG	2, 67, 74, 88	9
64	G	GG	42	

Table 2.1 Identification of 64 PCR amplification products cloned simultaneously in M13 mp18 and mp19. The ssDNA preparations which were subsequently analysed by DNA sequencing are underlined.

	34	58	69	146	186	199	221	263	279	297	319	320	328	331	350	379	381	404	410	422	428	439	463	470	475	506	512	522	531	532	533	536	570	
GB Ref	A	T	T	T	G	T	A	A	T	T	–	T	T	T	T	T	A	T	A	T	G	A	A	T	G	T	T	A	A	T	T	G	T	
PCR Sample																																		**Sequence Haplotype†**
1	•	•	•	C	•	•	•	•	C	C	AT	A	•	C	•	•	•	•	•	•	•	•	•	•	•	C	•	•	•	•	•	•	•	1C2
5	•	•	•	C	•	•	•	•	•	C	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
6	•	•	•	C	•	•	•	•	•	C	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	C	•	•	•	1A5
7	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
9	•	•	•	C	•	•	•	•	•	C	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	A	•	•	•	•	•	•	•	•	1B1
10	•	•	•	C	•	•	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1A1
11	•	•	•	•	A	•	•	•	•	•	•	A	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	..	•	•	•	•	•	1B5
14	•	•	•	C	•	•	•	•	•	C	AT	A	C	•	C	•	•	•	•	•	•	–	•	•	•	•	•	•	•	•	•	A	•	1B1
17	•	•	•	C	•	•	•	•	•	C	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1B1
21	•	•	•	C	•	•	•	•	•	C	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1B1
22	•	•	•	C	•	•	•	•	•	C	AT	A	•	•	•	•	G	•	•	•	•	•	G	•	•	•	•	•	•	•	•	•	•	1B1
23	•	•	•	C	•	•	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1A1
24	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	C	•	•	•	•	•	•	•	•	•	C	•	•	•	•	•	•	•	1C3
26	•	•	•	C	•	•	•	G	•	C	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1A2
28	•	•	•	C	•	•	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1A1
29	•	•	•	C	•	•	•	•	•	C	AT	A	•	•	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1A2
30	•	•	•	C	•	•	•	•	•	C	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1A2
32	•	•	•	•	•	•	•	•	•	•	•	A	•	•	•	C	•	•	•	•	A	•	•	•	•	C	•	•	•	•	•	•	•	
33	•	•	•	•	•	•	•	•	•	C	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	C	•	•	•	•	•	•	1B3
34	•	•	•	C	•	•	•	•	•	C	•	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
35	•	•	•	C	•	•	•	•	•	C	AT	A	•	•	•	•	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	1A2
36	•	•	•	C	•	•	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1B4
37	•	•	•	C	•	•	•	•	•	•	•	A	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
38	•	•	•	C	•	C	•	•	•	C	•	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
42	G	•	•	C	•	•	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
44	•	•	•	C	•	•	•	•	•	C	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	A	•	•	
47	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1A3
50	•	•	•	C	•	•	•	•	•	C	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1B1
51	•	•	•	C	•	•	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
54	•	•	•	C	•	•	•	•	•	C	AT	A	•	•	•	•	•	•	G	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1B1
55	•	•	•	C	•	•	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1A1
57	•	•	•	C	•	•	•	•	•	C	AT	A	•	•	•	•	•	•	•	•	•	•	•	C	•	•	•	•	•	•	•	C	•	1B1
58	•	•	•	C	•	•	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1A1
61	•	•	C	C	•	•	•	•	•	C	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
62	•	C	•	•	•	•	G	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	C	•	•	•	•	•	•	•	•	•	
63	•	•	•	C	•	•	•	•	•	C	•	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	G	•	•	•	•	

† Sequence haplotype designations are as given in Chapter 3. Haplotype identity was inferred by comparison with directly sequenced samples.

Figure 2.1 DNA sequence differences observed in 36 mp18 clones. "•" stands for identity with the GenBank consensus sequence, given at the top of the figure. "–" means deletion, or sequence not present. Boxed differences correspond to Taq misincorpation errors (determined by comparison with equivalent, directly sequenced, alleles). The shaded difference is a likely instance of phase disruption.

of a 52-bp long alternating purine-pyrimidine repeat (see consensus sequence, Figure A.1, Appendix A).

The 33 polymorphisms segregated as 27 sequence haplotypes. Twenty-four of the sequence haplotypes occurred only once, while 2 others were found at frequencies of 17% (sample numbers 10, 23, 36, 51, 55, 58 in Figure 2.1), and 17% (5, 17, 21, 26, 30, 50) in the sample.

Twenty-four of the 36 samples cloned were subsequently sequenced directly (discussed below) providing a means of distinguishing true polymorphisms (i.e. those present in the genomic DNA) from *Taq*-induced sequence errors 'captured' by the cloning of PCR-amplified material. Eighteen such *Taq* errors were observed among 12 of the 24 cloned sequences (boxed in Figure 2.1); the number of errors per sequence ranged from 1 to 4 (average 1.5). Sixteen (89%) of the errors were point mutations, of which 15 were transitions (9 T->C, 3 G->A, and 3 A->G) and 1 was a transversion. The other two differences were 1-bp deletions found at positions 439 and 522. Both deletions and 10 (62%) of the 16 point mutations (positions 253, 326, 331, 410, 463, 470, 475, 512, 536, 570) involved bases which fell in runs of 2 or more nucleotides of the same type (see Figure A.1). In addition to *Taq* misincorporation errors, a possible linkage phase disruption was observed in sequence 13, where the expected nucleotide at position 320, A, was replaced with a T (shaded in Figure 2.1). Polymorphic nucleotides flanking this position were otherwise linked as predicted by direct sequence analysis.

Another potential source of error was in the estimation of haplotype frequencies from cloned sequence data. Whereas direct sequence analysis of 48 chromosomes (both alleles of the 24 samples cloned) identified four major types of sequence haplotype, ranging in frequency from 42% (A-type haplotypes) to 32% (B), 18% (C), and 8% (D) (as described in Chapter 3), a different distribution of sequence haplotypes (46% A-, 46% B-, 8% C-, and no D-type haplotypes) was observed among the cloned sequences. Differences in the number and type of haplotypes could arise from random sampling effects (i.e. more A- and B-type alleles might have been isolated and sequenced by chance) or may reflect an underlying bias in the cloning efficiency of particular alleles.

In order to ensure that specific alleles were not being cloned preferentially, ten mp18 ssDNAs derived from a single PCR amplification product (sample 37, Table 2.1) were sequenced and compared. 200 bp, corresponding to positions 145 to 350 (-650 to -450 5' of the ß-globin gene), were examined. As shown in Figure 2.2, 50% of the cloned sequences derived from one allele and 50% from the other, suggesting that the distribution discrepancy observed for the larger sample was most likely a random effect.

Again, only 50% of the sequences analysed (3 from allele X and 2 from allele Y) were free from error: 4 *Taq* misincorporation errors and 2 linkage phase disruptions were observed among the other 5 sequences. *Taq* errors included 2 point mutations (A->G at 209, T->C at 213) as well as 2 length differences (1-bp deletion at 326, 2-bp insertion at 319). Phase errors were observed at position 145 in sequence 4, where the expected haplotype, T.T.-.A (Y allele), was replaced by C.T.-.A, and at position 319 in sequence 8, where C.C.AT.A (X allele), was replaced by C.C.-.A.

PCR Cloning as an Approach to Allele-Specific Sequence Analysis

The results presented above suggest that subsequent identification of simultaneously cloned PCR products *is* technically feasible, when specially-designed amplification primers are used for insert tagging: 97% of recombinants prepared as ssDNA were unambiguously identified, random selection of recombinant plaques resulted in the successful isolation of at least one copy of 91% of the PCR amplification products cloned, and 36 samples were retrieved in both orientations. However, although this approach eliminated much of the labour required for the purification, ligation, and transfection of individual samples, the method remained inefficient in terms of ssDNA preparation and sequencing. Given the retrieval rates observed here, nearly three times as many ssDNA templates would have to be prepared as are eventually analysed (more, if two or more copies of each sample are required in each orientation) and sample identification requires that all DNAs are sequenced (twice) before any subsequent examination.

Technical considerations aside, the quality of the data obtained from cloned PCR

	145	186	199	209	213	221	253	279	297	319	320	326	331	350	
GB Ref	T	G	T	A	T	A	A	T	T	--	T	T	T	T	
mp18 ssDNA															**Allele**
11	C	*	*	*	*	*	*	*	C	AT	A	*	*	*	X
31	C	*	*	*	*	*	*	*	C	ATAT	A	*	*	*	X
48	C	*	*	*	*	*	*	*	C	AT	A	*	*	*	X
59	C	*	*	*	*	*	*	*	C	AT	A	*	*	*	X
72	C	*	*	*	*	*	*	*	C	*	A	*	*	*	X
33	*	*	*	*	C	*	*	*	*	*	A	--	*	*	Y
44	C	*	*	*	*	*	*	*	*	*	A	*	*	*	Y
57	*	*	*	G	*	*	*	*	*	*	A	*	*	*	Y
84	*	*	*	*	*	*	*	*	*	*	A	*	*	*	Y
91	*	*	*	*	*	*	*	*	*	*	A	*	*	*	Y

Figure 2.2 Sequence comparison of mp18 clones derived from the a single PCR-amplified sample (no. 37, Table 2.1). Boxed differences represent Taq errors. Shaded positions are nucleotide sites affected by phase disruption.

products begs serious questions about the suitability of the method for population studies (Kwiatowski et al., 1991; Clark & Whittam, 1992). It is likely, for instance, (although comparative data were available for only a subset of the 36 samples sequenced) that 26 (79%) of the 33 variable positions identified in the cloned material were *Taq*-induced sequence artefacts. Where direct comparison was possible, 18 *Taq* errors (16 point mutations and 2 deletions) were identified, corresponding to an observed cumulative error frequency for point differences of 1.1×10^{-3} per nucleotide and for length differences of 1.4×10^{-4} (when the number of cycles of PCR amplification is taken into account, the estimated error rate per nucleotide synthesised, estimated according to Kwiatowski et al., 1991, is 6.9×10^{-5} for point mutations and 8.7×10^{-6} for length changes). Although comparatively low, this rate of error suggests that most 2.6 kb amplified molecules will contain one or more *Taq*-induced mistakes (Reiss et al., 1990). Therefore, at least two, and possibly three, copies of each cloned allele would have to be examined to ensure sequence accuracy.

Linkage phase disruption, another type of error identified among the cloned sequences, was less pervasive but, analytically, far more misleading. Phase disruption in cloned PCR products has been reported previously (Saiki et al., 1988; Shuldiner, Scott, & Roth, 1990; Jansen & Ledley, 1990a; 1990b; Kobayashi et al., 1990), and is thought to arise from excision repair of mismatched heteroduplex DNA (Jansen & Ledley, 1990a) formed during amplification of heterologous template (Nagamine, Chan, & Lau, 1989). Excision repair randomly corrects each heteroduplex mismatch so that the probability of correcting all sites to the original (true) haplotype is $(1/2)^{n-1}$, where n is the total number of mismatched sites in the molecule. Approximately five sites are expected to segregate, on average, between any two 2.6 kb alleles in a given sample (Chapter 3), therefore 15 out of 16 cloned heteroduplexes might be expected to experience some phase disruption. As approximately 50% of molecules are heteroduplexed following PCR amplification (Jansen & Ledley, 1990a), a significant proportion of cloned sequences are likely to suggest erroneous sequence haplotypes. Again, sequence analysis of several clones would be required to confirm the haplotype of any one allele.

M13 cloning is, in theory, an ideal way to prepare allele-specific ssDNA for dideoxy

chain termination sequencing. Several µgs of highly purified template can be generated from even a small quantity of genomic material. This template is easily analysed, gives clean, readable DNA sequence, and because the sequence derives from only one chromosome, the data are unambiguously interpretable in terms of the allelic sequence haplotype. In practice, however, cloning PCR products is difficult and time-consuming. The simultaneous cloning strategy eliminates much of the effort inherent to the approach, yet nevertheless remains inefficient, because many more clones must be processed than are ultimately sequenced. Most importantly, however, the sequences generated from cloned PCR products are error-prone and unreliable. Despite the relatively low misincorporation rate of *Taq* DNA polymerase, cloning captures rare errors in such a way that artefacts predominate in a collection of cloned sequences. In addition, frequent disruption of linkage phase in cloned heteroduplex DNA negates the principal advantage of cloning, the generation of allelic linkage data.

Direct Sequence Analysis of PCR-Amplified DNA

An alternative approach is to sequence PCR-amplified material directly. Direct sequencing circumvents many of the difficulties which plague sequence analysis of cloned material. Individual *Taq* misincorporation errors make up only a small fraction of the total product and are therefore not detected when PCR molecules are analysed collectively. Similarly, mismatch correction is not a problem because repair systems are not involved in the, entirely *in vitro*, process. Because the data reliably reflect sequence variation present in the genomic DNA, direct sequencing is also more efficient: amplification and analysis of one sequence per allele is typically all that is required.

A wide variety of methods have been devised which permit direct sequence analysis of PCR amplification products. Some of these methods involve sequencing denatured double-stranded molecules (Wrischnik et al., 1987; Wong et al., 1987; Engelke, Hoener, & Collins, 1988), another incorporates reverse transcription for single-stranded (RNA) sequencing (Stoflet et al., 1988), and still others use *Taq* (Innis et al., 1988) for 'cycle-

sequencing' under PCR-like conditions (Craxton, 1991; Lee, 1991). Invariably, however, the methods which produce sequence ladders of comparable quality to M13 involve the isolation of one of the two DNA strands prior to DNA sequence analysis. Such methods include *in vitro* synthesis of ssDNA, strand-specific enzymatic digestion, and physical isolation of ssDNA molecules.

In vitro amplification of ssDNA, also known as asymmetric PCR (Gyllensten & Erlich, 1988), is achieved using an unequal (or asymmetric) ratio of oligonucleotide primers during PCR amplification. Other methods for the production of ssDNA from double-stranded PCR molecules have a degradative, rather than synthetic, basis. One commonly used method involves I exonuclease digestion of the phosphorylated strand of a kinased PCR product (Higuchi & Ochman, 1989), while a related approach uses *E. coli* exonuclease III for strand-specific digestion (Ward, et al., 1989). Single-stranded DNA which is physically isolated from double-stranded PCR products, however, generally produces the highest quality DNA sequence. Single-stranded templates of 500 bp or longer can be isolated from agarose strand-separating gels (Sambrook, Fritsch, & Maniatis, 1989). A different method of ssDNA isolation utilises excess single-stranded M13 DNA to bind to one strand of the double-stranded PCR product, thereby preventing reannealing during the sequence reaction (Gal & Hohn, 1990). High affinity streptavidin-biotin binding is used for several other ssDNA purification methods. In each, PCR products are amplified with one normal primer and one biotinylated primer. Once amplified, the dsDNA is denatured and biotinylated strands separated from non-biotinylated strands by streptavidin-agarose column purification or by binding to streptavidin-coated magnetic beads (Hultman et al., 1989).

Solid Phase Sequencing of ß-Globin PCR Products

The least cumbersome and most reliable means of isolating ssDNA from PCR products was the streptavidin-biotin approach, using magnetic beads for solid support. This method, which has been adapted to the analysis of many genomic regions, was first applied to human ß-globin gene sequencing by Thein & Hinton (1991).

Single-stranded ß-globin template DNA was prepared using streptavidin-coated magnetic beads (Dynal UK Ltd.), according to the manufacturer's recommended protocol (Appendix A). The only specialised equipment required was a Magnetic Particle Concentrator (Dynal), a simple six-well magnetised eppendorf rack which immobilised beads during supernatant removal. New versions of the standard 5' and 3' amplification primers were also synthesised, so that a biotin molecule could be directly incorporated at the 5' end of each oligonucleotide (Appendix A). These biotinylated primers were stored and used like normal amplification primers. PCR amplification, using 10 pmol each of one normal and one biotinylated primer, was performed as described in Appendix A. Lower total primer concentration does not affect product amplification but does reduce the number of competing (i.e. non-extended) biotinylated primers present in the final reaction mixture. Following amplification, biotinylated PCR products were bound to the streptavidin-coated beads, washed, and alkali denatured. The non-biotinylated strand, and all other contaminants, were removed from the immobilised (biotinylated) strand. This purified ssDNA was then resuspended in distilled water and sequenced.

In general, 50 µl of the 3 kb PCR product (50 ng/µl) was sufficient for one sequencing reaction (this slightly larger PCR product, whose amplification is described in Appendix A, replaced the 2.6 kb fragment used in the cloning investigation). DNA sequencing was the same as for M13 template, except that labelling was performed on ice to enhance sequence close to the primer (R. Olds, personal communication). Although the manufacturer recommended the removal of the biotinylated (i.e. template) strand from the sequence reaction prior to electrophoresis, no sequence distortion was observed when the denatured bead-sequence mixture was loaded directly to the gel (also reported by Gyllensten & Allen, 1993). Clean and consistent DNA sequence, 250 to 500 bp in length, was obtained routinely using template prepared with magnetic beads. The resulting sequence ladder was sometimes fainter than comparable M13-generated sequence, due probably to the smaller quantities of template involved, otherwise there was no detectable difference in sequence quality. A given sample always produced the same DNA sequence: no *Taq* errors were observed in any directly sequenced PCR product.

Single-stranded DNA prepared using streptavidin-coated magnetic beads derived, of course, from both alleles present in the sample. The sequence generated was, therefore, a composite of both alleles, with heterozygous positions identified by the presence of two termination fragments, each of half the normal band intensity, migrating to the same location in the gel. If a sample was homozygous, or heterozygous at only one site, the sequence of both alleles could be determined simultaneously. When multiple heterozygous sites were observed in the same sequence, however, the linkage phase of sites was impossible to distinguish. The large number of sequence differences present at the human ß-globin locus (Chapter 3) meant that such linkage ambiguity was a common feature in many of the samples analysed.

Allele-Specific Sequence Analysis

Difficulties with linkage phase ambiguity mean that direct sequence analysis of PCR-amplified DNA, although in many respects preferable to cloning, is not immediately applicable to the analysis of *allelic* sequence diversity in nuclear DNA. In order to determine the sequence of particular alleles, a method which distinguishes, or otherwise isolates, DNA amplified from specific chromosomes is required. Sequence haplotype data can then be derived directly from the resulting allele-specific DNA sequence.

There are several different ways in which allele-specific DNA sequence may be obtained. A heterogeneous collection of template molecules (derived from both alleles) can be sequenced using a primer specific to one allele, so that only sequence from the primed allele results (Gyllensten & Erlich, 1988). Such allele-specific sequence priming relies on the presence of multiple sequence differences in a relatively small region (the primer binding site), and is therefore highly sequence-dependent. Often, an easier option is to isolate allele-specific DNA first (by electrophoretic separation using Denaturing Gradient Gel Electrophoresis [DGGE, Myers et al., 1985], or allele-specific restriction enzyme digestion [Scharf, Long, & Erlich, 1988], for instance), then sequence the template using standard primers and conditions.

Other methods allow allele-specific DNA to be generated directly during PCR amplification. One such approach relies on dilution of the starting genomic template, so that DNA from only one molecule is subsequently amplified (Jeffreys, Neumann, & Wilson, 1990; Ruano, Kidd, & Stephens, 1990), while another method uses allele-specific amplification primers for the generation of allele-specific DNA (Wu et al., 1989; Newton, et al., 1989).

Polymorphic differences present at the human ß-globin locus were not amenable to allele-specific DNA sequence priming or restriction enzyme digestion. Nor were the comparatively labour-intensive DGGE and single-molecule dilution approaches appropriate for the analysis of large number of chromosomes. Therefore, the allele-specific PCR amplification method, also known as the Amplification Refractory Mutation System or ARMS (Newton et al., 1989), was adopted. A principal advantage of the technique was that once allele-specific amplification was achieved, single-stranded template preparation and sequence analysis was identical to that performed on diploid samples.

ARMS (i.e. allele-specific) primers were chosen after direct sequence analysis identified the nature and location of heterozygous positions in each sample. The sites favoured for ARMS amplification were those with the highest observed heterozygosity, to enable a single ARMS primer to be applied to the analysis of many different samples. Polymorphic positions at either end of the 3 kb ß-globin fragment were targeted preferentially so that, when used in conjunction with one of the standard biotinylated amplification primers, as great a proportion of the original PCR product as possible could be specifically amplified and isolated. In many cases, all heterozygous sequence positions were encompassed within a single ARMS product. Other samples were amplified specifically twice, using different ARMS primers, to obtain allele-specific sequence for all ambiguous sites.

Once the ideal targets for amplification were identified, allele-specific primers were chosen and synthesised. Each ARMS primer was designed to anneal to the appropriate strand, with its 3'-most base complementary to the site to be amplified. In addition, a

single deliberate mismatch was placed at position -3 (three nucleotides 5' of the allele-specific base) to destabilise primer-template annealing at the 3' end of the oligonucleotide (Newton et al., 1989; Kwok et al., 1990) and maximise primer specificity. A single primer, specific to only one of the two bases present at a polymorphic site, was tested initially. Once optimised, the 'antiARMS' primer (i.e. the oligonucleotide used to preferentially amplify the other allele present at the polymorphic site) was also synthesised and tested.

Each primer was tested at its optimal annealing temperature (as predicted by computer analysis, Appendix A) with both positive (homozygous for the amplified base) and negative (homozygous for the non-amplified base) controls; amplification specificity was confirmed by the absence of product in the negative control. The annealing temperature was increased, as required, to ensure allele-specific amplification (PCR reaction conditions are given in Appendix A). Following optimisation, each ARMS primer was used to amplify allele-specific DNA from the relevant samples for direct sequence analysis.

In the course of the sequence investigations described in this thesis, 14 ARMS primers for 6 sites (positions 145, 917, 1358, 2636, 2792, 2945 in Figure 3.2) were designed, synthesised, and tested. Eleven of these primers (Tables A.2 and A.5, Appendix A) were subsequently used for allele-specific amplification. The primers which did not work either did not anneal specifically or would not amplify a single allele. In one instance (position 2945), allele-specific template was amplified from G, but the corresponding antiARMS primer could not amplify the T nucleotide specifically. While the presence of an additional internal mismatch helps, in many cases, to destabilise otherwise unrefractory sequence differences, primer destabilisation is not always sufficient, and other factors, perhaps intrinsic to *Taq* polymerase itself (Huang, Arnheim, & Goodman, 1992) must also affect single nucleotide discrimination. In contrast, the total size of the amplified product appears to have little effect on primer specificity. Allele-specific PCR products, ranging from 500 bp to 2.9 kb, were amplified with equal integrity.

Allele-specific ssDNA was isolated and sequenced in precisely the same manner as PCR

product amplified normally. The resulting DNA sequence (Figure 2.3) was unambiguous, revealing the relationship of polymorphic sites in the single amplified allele. The sequence haplotype of the other allele was most often derived by comparison with the original, diploid, sequence. In other instances, both alleles were specifically amplified and analysed. DNA sequence analysis of multiple amplification products from the same sample provided several opportunities to confirm observed differences and ensure sequence accuracy. Allele-specific DNA sequences always coincided with the compound sequence identified in a heterozygous sample.

Benefits and Drawbacks of Direct Sequence Analysis

Direct sequence analysis of PCR-amplified ß-globin alleles was, in most respects, preferable to cloning and sequencing. Single-stranded DNA was simply and rapidly isolated using streptavidin-coated magnetic beads. Using minimal equipment and reagents, the magnetic bead method produced single-stranded DNA from a single PCR product in 30 minutes (12 samples could be prepared simultaneously in less than 2 hours). In contrast, the M13 cloning protocol required 1 day for insert preparation, 1 day for ligation/transfection, and 1 to 4 days for ssDNA preparation, depending on the number of recombinant plaques isolated. In addition, with direct analysis there was no need for time-consuming preliminary DNA sequencing, such as that required for clone identification.

The resulting DNA sequence was of comparable quality to M13, with low non-specific background and long base reads. In addition, the impact of *Taq* errors was ameliorated by the collective analysis of multiple amplified molecules, and the formation of heteroduplexed DNA (which in cloning can lead to linkage phase disruption) was inconsequential to template preparation. Direct sequence data therefore accurately reflected differences present at the genomic level. Finally, because direct sequencing simultaneously analyses DNA derived from both alleles, the method was more efficient than cloning at detecting allelic variation present in genomic DNA from a given individual.

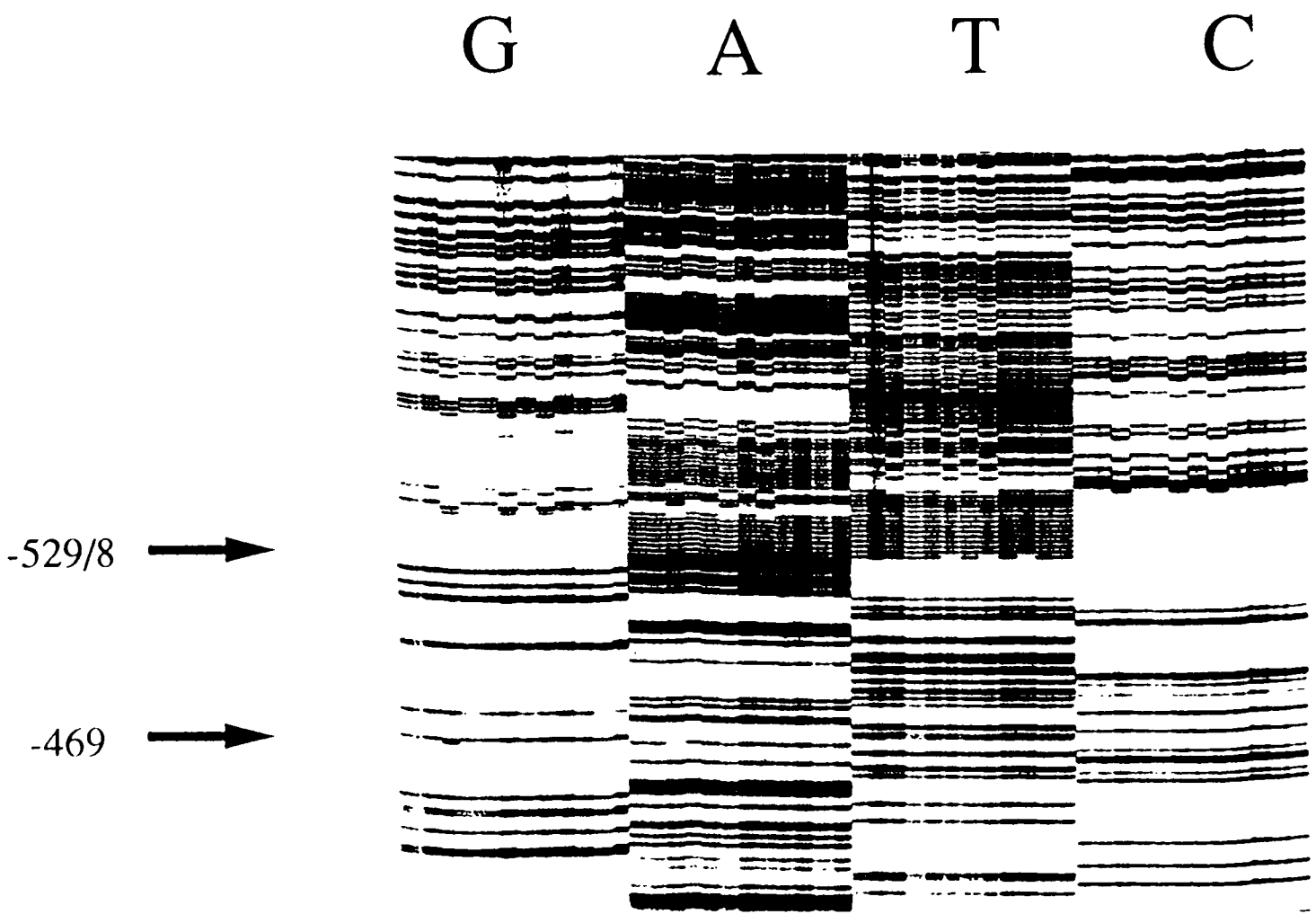


Figure 2.3 DNA sequence of 12 allele-specific templates amplified by the ARMS (Newton et al., 1989) method. Termination reactions were loaded in parallel and from left to right are ddGTP (G) 1-12, ddATP (A) 1-12, dTTP (T) 1-12, and ddCTP (C) 1-12. The arrows indicate length and point polymorphisms -529/-530 (positions 319/320, Figure 2.1) and -469 (position 379) bp 5' of the gene. In heterozygous samples the region upstream of the AT-repeat is uninterpretable because of length differences associated with the insertion polymorphism.

The principal drawback in using direct sequence analysis to investigate allelic sequence diversity at a nuclear locus like ß-globin arose from the inability to determine haplotype relationships when the sequenced template was heterozygous at multiple polymorphic sites (a common occurrence in the samples investigated here). Direct sequencing of allele-specific DNA, however, which was easily amplified using allele-specific PCR primers, circumvented these difficulties and yielded unambiguous DNA sequence. Therefore, a maximum of only three sequences per sample (one diploid, and one from each allele) was required for complete sequence haplotype resolution. In combination, diploid and allele-specific direct DNA sequencing accurately and efficiently identified all of the allelic sequence diversity present in the samples surveyed (Chapter 3).

The only real disadvantage of the direct sequence approach was that it tended to require considerably larger amounts of PCR product, and hence starting genomic template, than cloning did. This was particularly true in the analysis of the relatively large 3 kb ß-globin gene region, whose amplification was highly template-dependent. In addition, although much more efficient than cloning, direct sequence analysis was still relatively complex and time-consuming. It was for this reason that several indirect approaches to the investigation of genetic variation, which might provide the same information about ß-globin diversity as DNA sequence analysis but require much less effort, were also explored.

Indirect Detection of DNA Sequence Variation

Indirect methods of DNA polymorphism detection can be divided into two categories: those with the potential for identifying previously undescribed variable positions ('scanning' techniques) and those designed to detect the presence or absence of known polymorphisms ('diagnostic' methods). As the aim of analysis was the identification of *all* variation present in the sequences investigated, whether previously described or unknown, only scanning-type methods of mutation detection were considered. A number of such methods, utilising either restriction enzyme mapping, altered

electrophoretic mobility of ssDNA, or heteroduplex DNA analysis exist (reviewed in Rossiter & Caskey, 1990; Dianzani et al., 1993; Cotton, 1993; and Grompe, 1993). Each has its own strengths and weaknesses, so that the particular approach chosen depends upon the size and structure of the locus under investigation and the nature of the polymorphism expected, as well as the ultimate experimental aim.

In order to compete with DNA sequencing as an alternative approach to the analysis of ß-globin diversity, an indirect detection method would also have to use PCR-amplified DNA as its starting material, be able to screen a genomic region 3 kb in length, be highly efficient at detecting DNA sequence variation (particularly allelic polymorphism), and be relatively easy to implement and apply.

Two indirect methods of mutation detection, which between them incorporated each of the three major classes of approaches available, were tested. The application of these methods to the analysis of the 2.6 kb human ß-globin fragment, including a description of the protocol followed, the results obtained, and the relationship of the results to subsequently identified sequence variation, is given here (for further methodological details, see Appendix B). The viability of applying such indirect detection methods to the analysis of ß-globin polymorphism is also discussed.

Single-Strand Conformation Polymorphism Analysis of Four-Cutter Digested DNA

The well-known behaviour of single-stranded nucleic acids upon non-denaturing polyacrylamide gel electrophoresis, exploited for the isolation of single-stranded DNA (Maxam & Gilbert, 1980), was first applied to the detection of nucleotide polymorphisms by Orita et al. (1989a). Because single-stranded DNAs form semi-stable sequence-dependent conformations which affect fragment migration during electrophoresis, Orita et al. reasoned that single base changes would cause detectable differences in DNA mobility. This prediction was borne out, and the variable band patterns observed were named single-strand conformation polymorphisms (SSCPs) (Orita et al., 1989a; 1989b).

In their original report, Orita et al. (1989a) applied SSCP analysis to total genomic DNA, which was digested with a four-nucleotide-recognising restriction enzyme, denatured, and electrophoresed on a non-denaturing polyacrylamide gel. The blotted gel was then probed with labelled DNA or RNA (3 kb in length) to reveal a ladder of restriction fragment length polymorphism (RFLP) fragments, some of which showed SSCP variation. This strategy was rapidly supplanted by a PCR-based approach (Orita et al., 1989b) in which the fragment to be analysed (200 to 400 bp in length) was simultaneously amplified and labelled using end-labelled primers (Hayashi et al., 1989), denatured, and then electrophoresed under non-denaturing conditions.

As SSCP detection is highest for DNA fragments 400 bp or smaller (Orita, Sekiya, & Hayashi, 1990; Hayashi, 1991), the human ß-globin gene PCR product needed to be divided into smaller subregions for analysis. Separate amplification of six 450 bp fragments, requiring the synthesis and optimisation of ten additional amplification primers, was not considered a feasible option. Instead, as described for genomic DNA in the original SSCP report, the 2.6 kb PCR product was digested with four-cutter restriction enzymes to generate smaller fragments suitable for simultaneous SSCP analysis. By electrophoresing non-denatured and denatured aliquots of each digest on the same gel (Figure 2.4), both RFLP and SSCP detection methods could be assessed.

Twelve 2.6 kb ß-globin gene fragments, derived from samples from Espiritu Santo, Vanuatu, were simultaneously amplified and labelled by the incorporation of radioactively-labelled dCTP during the PCR reaction (a variation of the strategy suggested by Hayashi et al., 1989). Equal aliquots of each labelled amplification product were then digested with one of five four-cutter restriction enzymes, *Sau* 3A I, *Dde* I, *Hinf* I, *Alu* I, or *Hae* III, chosen for the size of the restriction fragments generated (as predicted by MacVector sequence analysis [International Biotechnologies, Inc.] of the GenBank consensus sequence). Digested PCR products were electrophoresed through non-denaturing polyacrylamide gels containing glycerol (moderate to high concentrations of glycerol enhance SSCP resolution). One half of each digest was electrophoresed normally (i.e. non-denatured) and the resulting ladder of fragments (i.e. the left side of the autoradiograph shown in Figure 2.4) scored for RFLP polymorphisms

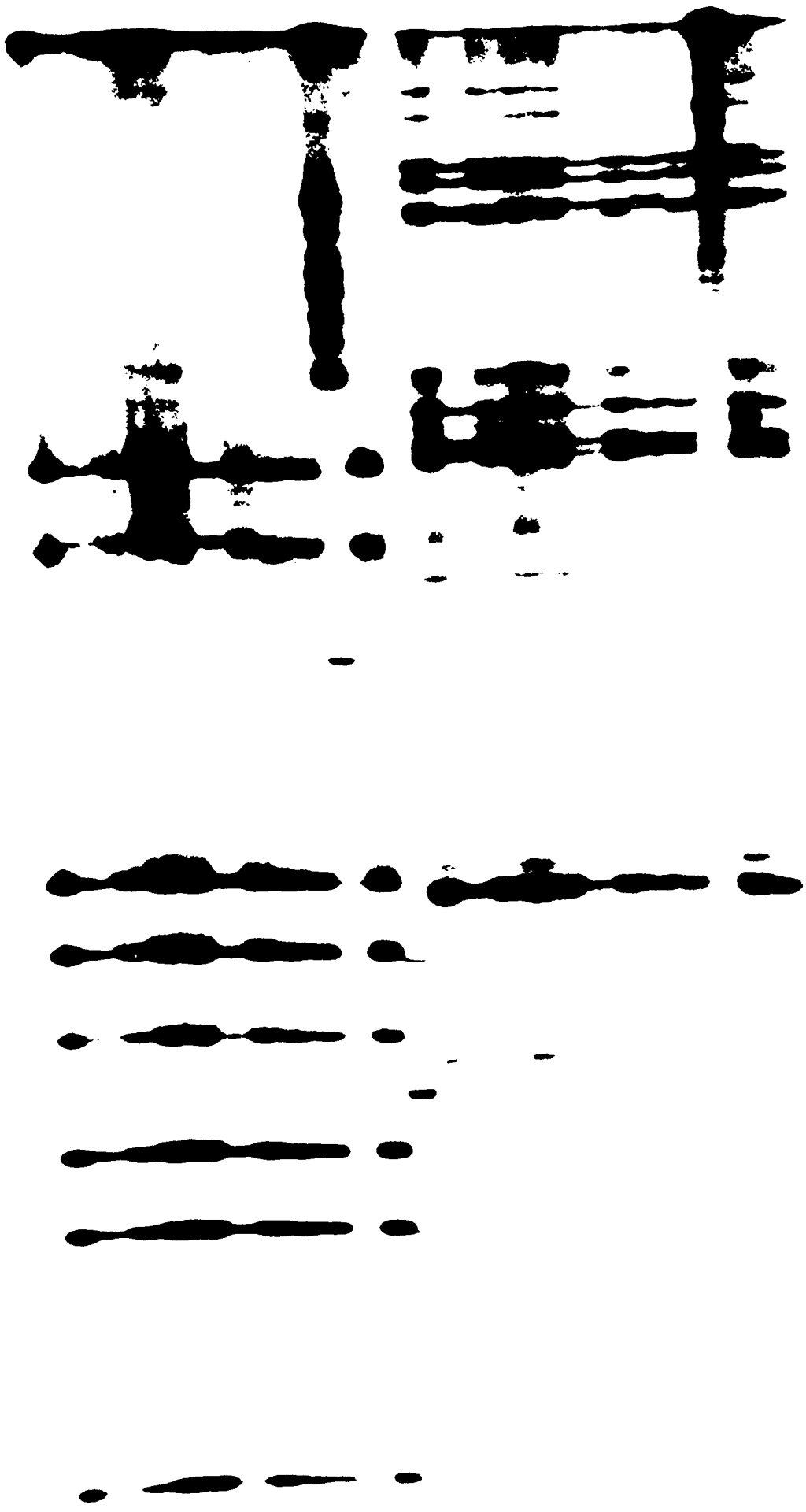


Figure 2.4 Combined RFLP-SSCP analysis of 12 samples from Espiritu Santo, Vanuatu. In this auto-radiograph the samples have been digested with Hinf I enzyme. The left-most lanes show the normal electrophoresis pattern for the non-denatured samples (1-12, left to right). The right-most lanes show the SSCP pattern for the same samples - aliquots denatured with formamide prior to electrophoresis. The central lane is φX174 Hae III-digested DNA.

resulting either from restriction enzyme site modification or length differences. The other half of each digested sample was denatured with formamide prior to electrophoresis and the resulting ssDNA bands (right side, Figure 2.4) scored for the presence or absence of SSCP mobility shifts.

Combined RFLP-SSCP analysis of eleven samples (the results for one sample, 10 in Figure 2.4, were disregarded due to problems with its electrophoresis) detected length and SSC polymorphism in 21 of the 45 electrophoretically detectable fragments (Table 2.2). The size of restriction fragments generated by each enzyme corresponded to those predicted: no polymorphisms were found which resulted in the gain or loss of a restriction enzyme recognition site. Minor length differences, involving DNA fragments which included the first 350 bp of the 2.6 kb PCR product (and therefore corresponding to insertion/deletion polymorphisms observed in cloned and directly sequenced samples), were observed in all five restriction enzyme digests. In most cases, however, limited electrophoretic resolution meant it was impossible to determine the exact nature of the length difference, or to distinguish whether samples were homozygous or heterozygous for particular length changes.

SSCP analysis detected polymorphism in virtually all of the DNA fragments in which it was present (genotypes of 6 of the 11 samples surveyed were subsequently determined by DNA sequence analysis). The two exceptions were *Dde* I, fragment 1 (512 bp) and *Alu* I, fragment 2 (737 bp). Unlike other fragments of equivalent size in which SSCPs were detected, these each contained only one point polymorphism, a difference which may have been insufficient to generate a conformation effect.

In addition to detecting the simple presence or absence of polymorphism, SSCP analysis also revealed the existence of more than two alleles in 10 of the 21 variable fragments (Table 2.2). For the six samples which were sequenced, the number of alleles identified by SSCP analysis was compared to the number subsequently found to be present. Detection efficiency, calculated as the ratio of the number of alleles observed to the number expected, varied from 20 to 100%. When digest fragments were grouped according to length, average detection efficiency was found to be lowest in

ENZYME			RFLP POLYMORPHISM			SSCP POLYMORPHISM		
Fragment Number	Size (bp)	Base Positions	Site	Length Observed	Length Expected	Observed	Expected	Number of Alleles
	Sau 3A I							
1	1161	1481-2641	-	-	-	+	+	2
2	627	542-1168	-	-	-	+	+	2
3	442	1-442	-	+	+	+	+	3
4	156	1325-1480	-	-	-	+	+	3
5	156	1169-1324	-	-	-	-	-	1
6	99	443-541	-	-	-	+	+	2
	Dde I							
1	512	1622-2133	-	-	-	-	+	1
2	371	1-371	-	+	+	+	+	4
3	276	379-654	-	-	-	+	+	2
4	273	2369-2641	-	-	-	+	+	3
5	201	915-1115	-	-	-	-	-	1
6	185	2134-2318	-	-	-	-	-	1
7	180	735-914	-	-	-	+	+	3
8	147	1330-1476	-	-	-	+	+	4
9	145	1477-1621	-	-	-	-	-	1
10	89	1204-1292	-	-	-	-	-	1
11	88	1116-1203	-	-	-	-	-	1
12	50	2319-2368	-	-	-	-	-	1
13	45	690-734	-	-	-	-	-	1
	Hinf I							
1	790	1346-2135	-	-	-	+	+	2
2	568	1-568	-	+	+	+	+	3
3	341	569-909	-	-	-	+	+	2
4	284	2136-2419	-	-	-	-	-	1
5	222	2420-2641	-	-	-	+	+	2
6	188	1158-1345	-	-	-	-	-	1
7	148	910-1057	-	-	-	-	-	1
8	75	1083-1157	-	-	-	-	-	1
	Alu I							
1	760	1316-2075	-	-	-	+	+	2
2	737	564-1300	-	-	-	-	+	1
3	464	1-464	-	+	+	+	+	4
4	217	2323-2539	-	-	-	-	-	1
5	130	2193-2322	-	-	-	-	-	1
6	79	485-563	-	-	-	+	+	2
7	67	2540-2606	-	-	-	+	+	2
8	49	2087-2143	-	-	-	-	-	1
9	35	2144-2192	-	-	-	-	-	1
	Hae III							
1	897	1252-2148	-	-	-	+	+	3
2	640	1-640	-	+	+	+	+	3
3	272	980-1251	-	-	-	-	-	1
4	236	2406-2641	-	-	-	+	+	2
5	208	772-979	-	-	-	+	+	2
6	131	641-771	-	-	-	-	-	1
7	101	2305-2405	-	-	-	-	-	1
8	81	2224-2304	-	-	-	-	-	1
9	75	2149-2223	-	-	-	-	-	1

Table 2.2 Results of RFLP-SSCP analysis of the 2.6 kb human ß-globin gene region in 11 unrelated diploids from Espiritu Santo, Vanuatu. '+/-' refers to the presence/absence of polymorphism in each electrophoretically-detectable fragment.

the largest fragments and highest in the smallest fragments (Table 2.3). The observed trend towards greatest detection in smaller-sized fragments is consistent with other published reports (Hayashi, 1991).

Despite relatively high rates of allelic detection for many of the variable fragments, a direct comparison between sequence haplotype and SSCP polymorphism revealed surprisingly few correlations between allelic variants identified by SSCP analysis and the underlying genotype (Figure 2.5). The presence of multiple variable sites in large (i.e. low SSCP detection) DNA fragments may partially explain this unexpected observation. Alternatively, errors in autoradiograph interpretation, which was particularly difficult in the case of larger fragments, could also be to blame.

RFLP-SSCP analysis of the ß-globin PCR product was successful in detecting much of the polymorphic variation present at the DNA sequence level. Restriction enzyme digestion produced a large number of overlapping DNA fragments which demonstrated both length polymorphism and ssDNA mobility shifts, and allelic detection efficiency for those fragments displaying SSC polymorphism averaged 70%. The approach was by no means definitive, however. Allelic detection efficiency was shown to vary inversely with fragment length, averaging less than 50% for fragments 500 bp or longer. As one-fifth of the DNA fragments generated by the five four-cutter digests fell within this size-range, mutation detection was severely limited for a significant subset of the fragments generated. No matter what size fragments were produced, however, results suggest that it is unlikely that all sequence polymorphisms present in a single sample would be detected by the two methods or that resultant SSCPs could be related in any meaningful way to the underlying sequence haplotype.

Taq Polymerase Extension of Carbodiimide-Modified Heteroduplex DNA

Another mutation detection method tested was the PCR-based carbodiimide (CDI) method of Ganguly and Prockop (1990). At the time the experimental work was undertaken, this was considered the simplest of the available heteroduplex-based detection methods (i.e. those in which heterozygous template is denatured and

Size Range	Total Number of Fragments	Percent Detection
>1000	1	50
750-1000	4	42.5
500-750	4	53.1
250-500	8	76
100-250	15	90
<100	12	100

Table 2.3 SSCP detection efficiency (the number of alleles observed relative to the number expected), averaged over fragments of similar size. Estimates are based on results from the 6 samples for which DNA sequence data were available.

Sample No.		145	297	307	319	320	327	379	508	532	906	1358	1416	1423	2008	2554	2634	2636
1	*1B5*	T	T	T	–	A	C	C	T	T	C	C	T	C	T	G	G	C
	1C4	T	T	C	–	A	C	C	C	T	T	G	T	C	C	G	G	A
	Sau 3A I	1/1							1/1		1/2		1/2			1/1		
	Dde I	1/1							1/2		2/3		1/3			2/2		
	Hinf I	1/2									1/2		1/2			1/2		
	Alu I	1/1							1/2							1/1		
	Hae III	1/2									1/2		1/2			1/1		
5	*1B1*	C	C	T	AT	A	C	T	T	T	C	C	T	C	T	G	G	C
	1B5	T	T	T	–	A	C	C	T	T	C	C	T	C	T	G	G	C
	Sau 3A I	1/2							1/1		1/1		1/1			1/1		
	Dde I	1/4							1/1		1/1		1/1			1/2		
	Hinf I	1/1									1/1		1/1			1/1		
	Alu I	3/3							1/2							1/1		
	Hae III	2/2									2/2		1/1			1/1		
6	*1B1*	C	C	T	AT	A	C	T	T	T	C	C	T	C	T	G	G	C
	1C1	T	T	T	–	A	-	T	C	T	T	G	T	C	C	G	A	A
	Sau 3A I	2/2							1/1		1/1		1/1			1/1		
	Dde I	2/2							1/1		1/1		1/1			1/1		
	Hinf I	1/1									1/1		1/1			1/1		
	Alu I	3/3							1/2							1/1		
	Hae III	2/2									1/2		1/1			1/1		
7	*1A1*	C	C	T	–	T	C	T	T	T	C	C	G	C	T	G	G	C
	1B1	C	C	T	AT	A	C	T	T	T	C	C	T	C	T	G	G	C
	Sau 3A I	2/3							1/1		1/1		1/1			1/1		
	Dde I	2/3							1/1		1/1		1/1			1/1		
	Hinf I	1/2									1/1		1/1			1/1		
	Alu I	2/3							1/2							1/1		
	Hae III	2/2									2/2		1/2			1/1		
9	*1A1*	C	C	T	–	T	C	T	T	T	C	C	G	C	T	G	G	C
	1D1	T	T	T	–	T	C	T	C	T	T	G	T	T	C	C	G	A
	Sau 3A I	3/3							1/1		1/2		1/3			1/1		
	Dde I	2/3							1/1		1/2		1/3			1/1		
	Hinf I	1/2									1/1		1/2			1/1		
	Alu I	2/2							1/2							1/1		
	Hae III	2/2									1/2		1/2			1/1		
11	*1C2*	C	C	T	AT	A	C	T	C	T	T	C	T	C	C	G	G	A
	1D1	T	T	T	–	T	C	T	C	T	T	G	T	T	C	C	G	A
	Sau 3A I	1/1							2/2		2/2		1/3			2/2		
	Dde I	1/1							2/2		2/3		4/4			3/3		
	Hinf I	1/2									1/2		1/2			1/2		
	Alu I	3/4							1/1							1/2		
	Hae III	3/3									1/1		1/3			1/2		

Figure 2.5 RFLP-SSCP maps for the 6 samples for which DNA sequence data were available. The fragments generated by each restriction enzyme digest and associated SSCP polymorphism is shown. Each fragment block is defined by the sequence polymorphisms it encompasses (i.e. blocks are not drawn to scale). The numbers in each block are designations which refer to the alleles identified during SSCP analysis. Monomorphic fragments have been left blank. The allelic sequence haplotypes identifed for each sample (sequence haplotype designations as given in Chapter 3) are shown above each RFLP-SSCP map.

reannealed, and the resultant heteroduplexes subjected to mismatch cleavage or modification). It has, however, almost certainly been supplanted in this regard by more recent techniques involving the electrophoretic detection of DNA heteroduplexes in non-denaturing polyacrylamide gels (e.g. White et al., 1992; Wood, Tyfield, & Bidwell, 1993).

In the CDI method, heteroduplexed DNA molecules are treated with water-soluble CDI, a chemical which reacts with mismatched Gs and Ts. Primer extension with *Taq* polymerase, under conditions in which DNA synthesis is terminated at the modified base, is used to identify the site of mismatch in the CDI-modified DNA. Although the method can only detect a single mutation, i.e. the one nearest the PCR primer, it was reasoned that with multiple priming sites many, if not all, polymorphic sites within the 2.6 kb ß-globin fragment could be detected. For this reason, PCR primers which annealed to different strands, and at different positions in the ß-globin product, were investigated.

Twelve 2.6 kb PCR products were amplified, denatured by heating, and slowly reannealed to generate allelic heteroduplexes. A portion of each heteroduplexed sample was incubated with CDI solution for 3 hours (in siliconised tubes, to prevent adherence of CDI-modified DNA), extracted, and heated to destroy residual CDI. CDI-modified DNA was then used as template in a one-cycle PCR reaction, using a single oligonucleotide primer. Four primers, including the original 5' and 3' PCR amplification primers (AP 1 and AP 2) and two internal primers (AP 3 and AP 4), were tested. Similar amplification of non-CDI treated control DNA was also performed. Reaction products, labelled during amplification by direct incorporation of ^{32}P-labelled dCTP, were electrophoresed on denaturing polyacrylamide gels so that the length of each amplified ssDNA (hence, the position of the relevant mismatch) could be determined.

No single-stranded PCR amplification products were observed for any of the twelve CDI-modified DNA templates tested. Control (non-CDI-modified) DNAs did, however, result in isotopically-detectable full length products. The experiment was repeated twice, and

each time the same results were obtained. Subsequent DNA sequence analyses revealed that point polymorphisms, which would have resulted in either mismatched Ts, or Gs, *were* downstream of each PCR primer in the samples tested. The sizes of the amplification products expected were 145 bp for AP 1, 87 bp for AP 2, 128 bp for AP 3, and 73 bp for AP 4. No such fragments were observed.

The CDI method of mutation detection did not successfully identify any length or point polymorphism present in the ß-globin PCR product. Neither heteroduplex formation nor the priming strategy were to blame, however, because non-CDI treated PCR products were successfully amplified. Rather, it seems that CDI modification itself interfered with, or prevented, *Taq* polymerase primer extension. Significant template losses following CDI-modification were possible, due both to the adherence of CDI-treated DNA to plastic and to the altered solubility of CDI-modified DNA, which makes ethanol precipitation difficult (Ganguly and Prockop, 1990). Alternatively, the presence of residual CDI in the modified template might have inhibited *Taq* polymerase activity during PCR.

Further optimisation was not attempted because subsequent investigation, using DNA sequence analysis, revealed several clusters of closely-spaced polymorphic sites which could not be detected by the primer extension method. This drawback, combined with the methodological complexity of the approach (the CDI method was more difficult to implement than SSCP analysis and nearly as time-consuming as DNA sequencing), made the successful application of the method to the study of ß-globin genetic diversity highly unlikely. It is important to note that no cleavage-based method of mutation detection is particular well-suited to the analysis of *allelic* sequence diversity of genetic loci: although the position of polymorphic sites is identified, such methods provide no indication of the way in which variable sites are linked along the chromosome.

Suitability of Indirect Methods for the Investigation of ß-Globin Diversity

Indirect detection of genetic variation would be a viable alternative to DNA sequence analysis if the method or methods adopted were quicker and easier than sequencing,

identified all (or nearly all) variation present, and detected allelic polymorphism. The two approaches tested, which are representative of the indirect methods currently available, suggest that it is unlikely that any indirect method can compete with the comprehensive description of ß-globin diversity provided by DNA sequence analysis.

One of the two methods investigated here represented a quicker alternative for ß-globin analysis than DNA sequencing. The electrophoretically-based RFLP-SSCP analysis was relatively easy to implement and apply, and sample processing rapid, at least in comparison with direct sequence analysis of a similar number of samples. The CDI method, however, required a much larger number of manipulations, making it unfeasible for the analysis of a large number of samples, and was nearly as time-consuming as sequencing.

DNA sequence analysis, which detects all polymorphic variation, was better at identifying variable sites at the ß-globin locus than either of the indirect methods tested. RFLP analysis identified no point polymorphisms and provided only a rough indication of the region in which length differences were located. SSCP detection efficiency, which was greatest in smaller-sized restriction fragments, averaged less than 75% overall. The CDI method, if it had worked, probably would have detected 70 to 80% of the polymorphic sites present in the 12 samples.

Finally, neither of the indirect methods investigated was particularly well-suited to the analysis of *allelic* sequence diversity at the human ß-globin locus. The subdivision of the 2.6 kb PCR product, required for SSCP electrophoretic resolution, made it difficult, if not impossible, to interpret the observed patterns of SSCP variation in terms of the allelic polymorphism of the total ß-globin gene region. The CDI detection protocol, which was applicable to the whole ß-globin amplification product, only identified *where* the polymorphic sites lay in the amplified region, not *how* they were linked together.

In summary, cleavage-based heteroduplex methods, such as the CDI indirect detection method, are methodologically complex and produce data inappropriate for a survey of allelic sequence diversity. Electrophoretically-based detection methods like SSCP

analysis, although easier to implement than DNA sequencing, require the analysis of small genomic regions, and this negates their usefulness for the analysis of regions as large as the human ß-globin gene. Therefore, neither type of method represent a good alternative to DNA sequence analysis.

The Investigation of DNA Sequence Variation at a Human Nuclear Locus

A variety of methodological approaches to the assessment of allelic sequence diversity at the human ß-globin gene, a nuclear DNA locus, were investigated before a standard experimental protocol was adopted.

M13 cloning of PCR-amplified material was explored as one means of generating high quality, allele-specific, single-stranded DNA for sequencing. However, pilot analysis of 36 alleles, using a PCR-based sample labelling strategy for simultaneous cloning of multiple amplification products, was disappointing. The sequences of the cloned PCR products contained numerous *Taq* polymerase misincorporation errors and linkage phase disruptions. The method was also inherently inefficient, requiring the processing of many more clones than were ultimately sequenced.

An alternative to cloning, direct sequence analysis of PCR-amplified DNA, was also investigated. Single-stranded DNA preparation using streptavidin-coated magnetic beads, which was quicker and easier than M13 cloning, generated purified template directly from amplified material. The resulting high-quality DNA sequence derived from multiple amplified molecules, so that individual *Taq* errors were not detected. The sequences of both alleles present in a sample were also revealed simultaneously, making direct sequence analysis more efficient than cloning at identifying allelic variation. The linkage phase of multiply heterozygous sites was determined by sequencing allele-specific PCR amplification products.

Due to the time-consuming nature of DNA sequence analysis, two indirect methods for the detection of sequence polymorphism were explored as viable alternative

approaches. The RFLP-SSCP method was rapid to perform and detected much of the polymorphic variation present in a 2.6 kb long ß-globin fragment. The approach was plagued, however, by limited detection efficiency in a significant proportion of the restriction fragments generated, and a lack of correlation between SSCP alleles and underlying genotypes. The CDI method, in contrast, was not simple to implement and did not work. If it had worked, it would not have detected all the variable sites present at the ß-globin locus, nor could it have provided any indication of allelic relationship among the polymorphisms identified.

Thus, direct DNA sequence analysis, using allele-specific amplification for linkage determination, proved the most rapid and efficient means of investigating allelic sequence diversity at a human nuclear locus. The method was easier to apply than PCR cloning and resulted in much more reliable sequence data. In addition, as a means of investigating both nucleotide and allelic sequence diversity in a large genomic region, it compared favourably with the indirect methods tested. No indirect detection method could identify precisely which changes had occurred where, or indicate how those changes were related to one another, in an allelic sense, across the entire 3 kb region.

Direct sequence analysis of PCR-amplified DNA was therefore applied to the assessment of ß-globin sequence diversity in humans. In total, 173 chromosomes from five populations (Chapter 3) were analysed. The results of these investigations, and their implications for our understanding of the evolutionary processes acting at the human ß-globin locus, are discussed in the remainder of the thesis.

Chapter 3

DNA Sequence Variation at the Human ß-Globin Locus

The best way to identify the evolutionary processes which contribute to the maintenance of genetic variation in human populations is DNA sequence analysis of alleles present at multiple unlinked loci. Although there are numerous methods available for studying the nucleotide sequence of genomic regions, few are suited to population-based analysis of diploid DNA. As a consequence, surveys of human nuclear DNA diversity lag behind equivalent mtDNA studies and human population genetic investigation at the sequence level remains confined to a single genetic locus. Direct sequence analysis of allele-specific PCR amplification products, the approach adopted here (Chapter 2 and Appendix A), makes the detection of allelic sequence variation in heterozygous samples finally feasible. Armed with this new experimental strategy, the first survey of DNA sequence diversity at a human nuclear locus, the ß-globin gene, was undertaken.

DNA sequence diversity contains important information about the past, and on-going, impact of molecular, selective, and demographic processes in human populations. DNA polymorphisms begin as *de novo* mutations, generated by nucleotide substitution or by the insertion or deletion of genetic material. These mutations arise on a background of existing polymorphic alleles and the resulting linkage relationships (described in terms of allelic sequence haplotypes) reflect the relative timing of genetic changes and may also, if disrupted, indicate the influence of interallelic recombination. The frequencies of sequence haplotypes in populations are determined by rates of mutation and recombination, the extent of gene flow with other locales, the effects of natural selection, and the influence of random genetic drift. By evaluating observed genetic variation with respect to neutral theory expectations (Chapter 1), the relative contributions of these different evolutionary processes can be assessed.

Despite extensive previous investigation of the human ß-globin gene at the DNA sequence level, allelic sequence diversity at the locus has never been considered in light of the impact of such processes. Instead, the focus of molecular genetic analysis has been the identification of mutations which, when inherited in homozygous form, result in severe haemoglobinopathies such as sickle cell anaemia and ß thalassaemia; to date, well over 100 such mutations have been discovered in and adjacent to the ß-globin gene (Thein, 1993). These mutations occur in association with a smaller number of silent nucleotide differences, polymorphisms which have no apparent affect on gene functioning and which together define four common ß-globin gene 'frameworks' (Orkin & Kazazian Jr., 1984; Antonarakis, Kazazian Jr., & Orkin, 1985). For clinical scientists, these frameworks are the background upon which disease-causing mutations arise, and interesting only in terms of what they suggest about the origin and geographic distribution of haemoglobinopathies. Little attempt has been made to explore the relationship of the underlying sequence variation to larger evolutionary phenomena.

In order to investigate the impact of molecular, selective, and demographic influences on DNA sequence variation at the human ß-globin gene, the nucleotide and sequence haplotype polymorphism present in a collection of predominantly normal alleles was examined. In all, 173 chromosomes from five populations were sequenced and compared.* The genetic variation identified, and its relationship to previously detected polymorphism as well as theoretically-expected levels of sequence haplotype diversity, is discussed in this chapter.

The Data

173 chromosomes from five separate geographic locations (hereafter referred to as 'populations') were investigated for allelic sequence polymorphism in a 3 kilobase (kb)

* Twenty-three of the 173 sequences reported here were determined by J. Bond, Institute of Molecular Medicine, Oxford.

genomic region encompassing the human ß-globin locus (Figure 3.1). The sequence haplotypes identified, and their respective frequencies in each population, are given in Figure 3.2 (the consensus sequence, which derives from the empri:hshbb GenBank [Bilofsky & Burks, 1988] reference sequence, is given in full in Figure A.1, Appendix A).

In several of the populations surveyed, known heterozygotes for ß chain haemoglobinopathies (typed by protein electrophoresis) were preferentially chosen for sequence analysis. This strategy allowed variation among affected chromosomes to be evaluated, and aided in the process of inferring the sequence haplotype of normal alleles, as discussed in Chapter 1. As a consequence, however, the sample 'frequencies' of these sequence haplotypes were not representative of the underlying frequencies of affected alleles in each population. By taking into account previously reported gene and/or carrier frequencies,* an effective (as opposed to actual) number of sequence haplotypes was assigned to each mutant allele (i.e. 2A4, 4A4, 4C15, 5A11, and 5B22, Figure 3.2). These adjusted totals, which are given in parentheses in the figure, were then used in all subsequent analyses.

In the sample from Sumatra, two different sequence haplotypes with the Intron 1, nucleotide position 5 G->C $ß^{thal}$ mutation (position 994) were observed, corresponding to the reported occurrence of the mutation on two 3' RFLP haplotype backgrounds. As the relative occurrence of the two backgrounds in Indonesia is 2 to 1 (Lie-Injo et al., 1989), the frequency of each sequence haplotype was adjusted accordingly (the total frequency of the mutation in the sample, irrespective of background, remained 10%). The frequencies of the two $ß^{S}$ haplotypes in the Gambia were similarly scaled to reflect their observed 2 to 1 occurrence among the 22 sickle heterozygotes surveyed. Although care was taken to assure that the frequencies of mutant haplotypes were as close as possible to those reported previously, in practice, the relative occurrence of particular types of mutant alleles made little difference to estimates of nucleotide or haplotype diversity in the samples.

---

* Estimated frequencies of the affected chromosomes were taken as 10% for the Intron 1, nt 5 G->C $ß^{thal}$ mutation in Maewo (Hill et al., 1988), 10% for the Intron 1, nt 5 G->C $ß^{thal}$ mutation in Indonesia (A. S. M. Sofro, personal communication), and 7% for the $ß^{S}$ mutation in the Gambia (Hill et al., 1991).

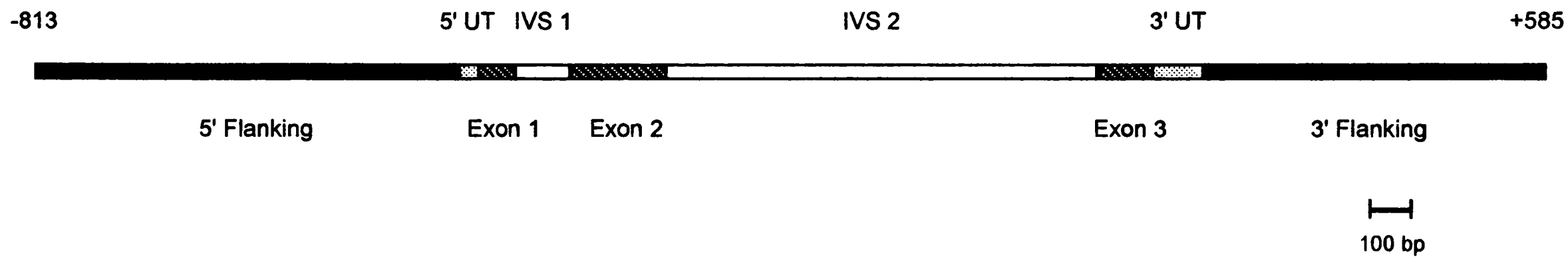


Figure 3.1 The 3.0 kb region sequenced, showing the Exons, Introns (abbreviated IVS for Intervening Sequence), and Untranslated (UT) regions of the human ß-globin gene.

	5' Flanking																			Exon 1		Intron 1	Intron 2					3' Flanking							
	130	138	145	284a	287	305	307	318	318a	320	322	325	325a	327	328	357	379	608	632	906	917	994	1358	1388	1446	1423	2068	2503	2654	2834	2836	2782	2924	2945	FREQ
GB Ref	G	T	T	–	T	C	T	A	–	T	T	T	–	C	T	A	T	T	T	C	A (Bs)	G (Bsh)	C	T	G	C	T	G	G	G	C	A	T	G	
Espiritu Santo, Vanuatu																																			
1A1	•	•	C	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	10
1A2	•	•	C	•	C	•	•	•	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	5
1A5	•	•	C	•	C	•	•	•	AT	A	•	•	•	•	•	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1
1A3	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	2
1B1	•	•	C	•	C	•	•	•	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	9
1B3	•	•	•	•	C	•	•	•	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	2
1B2	•	•	C	•	C	•	•	•	AT	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	1
1B4	•	•	C	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	1
1B5	•	•	•	•	•	•	•	•	•	A	•	•	•	C	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	2
1B7	•	•	•	•	•	•	•	•	AT	A	•	•	•	–	–	•	•	C	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	1
1C3	•	•	•	•	•	•	•	•	•	A	•	•	•	•	•	•	C	C	•	T	•	•	G	•	T	•	C	•	•	•	A	•	•	T	6
1C4	•	•	•	•	•	•	C	•	•	A	•	•	•	•	•	•	C	C	•	T	•	•	G	•	T	•	C	•	•	•	A	•	•	T	1
1C1	•	•	•	•	•	•	•	•	•	A	•	•	•	–	•	•	•	C	•	T	•	•	G	•	T	•	C	•	•	A	A	•	•	T	1
1C2	•	•	C	•	C	•	•	•	AT	A	•	•	•	•	•	•	•	C	•	T	•	•	•	•	T	•	C	•	•	•	A	•	•	T	2
1D1	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	C	•	T	•	•	G	•	T	T	C	•	C	•	A	•	•	T	2
1D2	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	C	•	•	•	•	G	•	T	T	C	•	C	•	A	•	•	T	1
1D3	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	C	•	T	•	•	G	•	T	T	C	•	C	•	A	•	•	T	1
																																			n= 48
Maewo, Vanuatu																																			
2A1	•	•	C	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	2
2A2	•	•	C	•	C	•	•	•	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	3
2A3	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	2
2A4	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	12 (1)
2B1	•	•	C	•	C	•	•	•	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	2
2B2	•	•	C	•	C	•	•	•	AT	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	1
2C2	•	•	C	•	C	•	•	•	AT	A	•	•	•	•	•	•	•	C	•	T	•	•	•	•	T	•	C	•	•	•	A	•	•	T	1
2D1	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	C	•	T	•	•	G	•	T	T	C	•	C	•	A	•	•	T	1
																																			n= 24 (13)
Eastern Highlands, Papua New Guinea																																			
3A1	•	•	C	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1
3B1	•	•	C	•	C	•	•	•	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	6
3B3	•	•	•	•	C	•	•	•	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	4
3B2	•	•	C	•	C	•	•	•	AT	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	1
3B8	•	•	•	•	•	•	•	•	AT	A	•	•	•	–	–	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	1
3C1	•	•	•	•	•	•	•	•	•	A	•	•	•	–	•	•	•	C	•	T	•	•	G	•	T	•	C	•	•	A	A	•	•	T	3
3C5	•	•	•	•	•	•	•	•	•	•	•	•	•	–	•	•	•	C	•	T	•	•	G	•	T	•	C	•	•	A	A	•	•	T	1
3C6	•	•	•	•	•	•	•	•	AT	A	•	•	•	–	–	•	•	C	•	T	•	•	•	•	T	•	C	•	•	•	A	•	•	T	6
3C7	•	•	•	•	•	•	•	•	AT	A	A	•	•	–	–	•	•	C	•	T	•	•	•	•	T	•	C	•	•	•	A	•	•	T	1
																																			n= 24
Sumatra, Indonesia																																			
4A1	•	•	C	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1
4A3	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1
4A4	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	1 (1)
4B5	•	•	•	•	•	•	•	•	•	A	•	•	•	•	•	•	C	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	1
4B6	•	•	•	•	•	•	•	•	•	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	1
4B3	•	•	•	•	C	•	•	•	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	1
4B1	•	•	C	•	C	•	•	•	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	3
4B2	•	•	C	•	C	•	•	•	AT	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	1
4B4	•	•	C	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	1
4B9	A	•	C	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	1
4C8	•	•	•	•	•	•	C	•	•	•	•	•	•	•	•	•	C	C	•	T	•	•	G	•	T	•	C	•	•	•	A	•	•	T	4
4C9	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	C	C	•	T	•	•	G	•	T	•	C	•	•	•	A	•	•	T	3
4C10	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	C	C	•	•	•	•	G	•	T	•	C	•	•	•	A	•	•	T	1
4C3	•	•	•	•	•	•	•	•	•	A	•	•	•	•	•	•	C	C	•	T	•	•	G	•	T	•	C	•	•	•	A	•	•	T	1
4C11	•	•	•	•	•	•	•	•	AT	•	•	•	•	•	•	•	C	C	•	T	•	•	G	•	T	•	C	•	•	•	A	•	•	T	1
4C12	•	•	C	•	C	•	•	•	AT	A	•	•	•	•	•	•	•	C	•	T	•	•	G	•	T	•	C	•	•	•	A	•	•	T	1
4C13	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	C	•	T	•	•	G	•	T	•	C	•	•	A	A	•	•	T	1
4C14	•	•	C	•	•	•	•	•	•	A	•	•	•	–	•	•	•	C	•	T	•	•	G	•	T	•	C	•	•	A	A	•	•	T	1
4C1	•	•	•	•	•	•	•	•	•	A	•	•	•	–	•	•	•	C	•	T	•	•	G	•	T	•	C	•	•	A	A	•	•	T	1
4C15	•	•	•	•	•	•	•	•	•	A	•	•	•	–	•	•	•	C	•	T	•	C	G	•	T	•	C	•	•	A	A	•	•	T	1 (2)
																																			n= 27 (28)
Gambia, Africa																																			
5A6	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	2
5A7	•	•	•	•	•	•	•	•	•	A	•	•	T	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1
5A8	•	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	2
5A9	•	•	•	•	C	•	•	T	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	A	•	•	•	•	•	•	2
5A10	•	•	C	•	C	•	•	T	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	1
5A11	•	G	•	•	•	•	•	•	AT	•	•	•	•	–	–	•	•	•	•	•	T	•	•	•	•	•	•	•	•	•	•	•	•	•	15 (2)
5B6	•	•	•	•	•	•	•	•	•	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	3
5B1	•	•	C	•	C	•	•	•	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	2
5B11	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	C	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	T	•	T	1
5B12	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	C	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	•	C	T	1
5B13	•	•	C	•	•	T	•	•	•	•	•	•	•	•	•	C	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	•	C	T	1
5B14	•	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	C	T	•	•	•	•	•	•	•	C	T	1
5B15	•	•	•	•	•	T	•	•	•	A	•	–	•	•	•	C	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	•	C	T	1
5B16	•	•	•	•	•	T	•	•	•	A	•	–	•	•	•	C	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	•	•	T	1
5B17	•	•	•	AT	•	T	•	•	•	•	•	–	•	•	•	C	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	•	•	T	1
5B18	•	•	C	•	•	T	•	•	•	•	•	–	•	•	•	C	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	•	•	T	1
5B19	•	•	C	•	•	T	•	•	•	•	•	•	•	•	•	C	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	•	•	T	1
5B20	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	C	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	•	•	T	1
5B21	•	•	•	•	•	•	•	•	•	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	•	•	T	2
5B22	•	•	•	•	•	T	•	•	•	A	•	•	•	•	•	C	•	•	•	•	T	•	•	•	T	•	•	•	•	•	•	•	•	T	7 (1)
5C18	•	•	•	•	•	T	•	•	•	A	•	–	•	•	•	C	•	•	•	T	•	•	G	•	T	•	C	•	•	•	A	•	•	T	1
5C19	•	•	•	•	•	•	•	•	•	A	•	•	T	•	•	•	C	C	•	T	•	•	G	•	T	•	C	•	•	•	A	•	•	T	1
5D4	•	•	•	•	C	•	•	T	•	•	•	•	•	•	•	•	•	C	•	T	•	•	G	•	T	T	C	•	•	•	A	•	•	T	1
																																			n= 50 (31)

Figure 3.2 ß-globin allelic sequence diversity observed in 87 diploid individuals from five human populations. Haplotype designations are coded to reflect the population of origin (numerical prefix), the type of sequence haplotype (A, B, C, or D), and identity with other sequences (e.g. all A1s, irrespective of the population they are found in, possess the same sequence haplotype). Adjusted frequencies (see text) are given in parentheses.

Nucleotide Diversity at the Human ß-Globin Locus

Thirty-four nucleotide positions were observed to vary in the sequences surveyed. Most of these differences consisted of simple, single nucleotide substitutions. Other differences involved the addition or deletion of one or more bases in the sequence. The nature of these variable sites is described here.

Nucleotide Polymorphism

Twenty-eight dimorphic nucleotide positions were identified among the 77 sequence haplotypes shown in Figure 3.2: 13 in the 5' flanking DNA (814 basepairs [bp] total length), 2 in Exon 1 (93 bp long), 1 in Intron 1 (130 bp) and 5 in Intron 2 (850 bp) of the ß-globin gene, and 7 in the 3' flanking region (582 bp). No polymorphic sites were observed in the 5' and 3' untranslated regions of the gene, or Exons 2 and 3.

Two of the 28 observed polymorphisms are known to affect the structure or synthesis of the ß globin protein chain. The A->T change at position 17 of Exon 1 (917 in Figure 3.2) is the mutation which leads to sickle cell anaemia. This non-synonymous substitution results in the replacement of the amino acid glutamic acid with valine, at residue 6 of the ß-globin polypeptide chain (Ingram, 1957). The polymorphism at 994 causes a RNA processing defect which results in ß thalassaemia (Treisman, Orkin, & Maniatis, 1983). Thus, although the latter difference occurs in non-coding DNA (Intron 1), both polymorphisms are non-silent nucleotide changes.

Four of the 28 nucleotide polymorphisms were observed only once in 144 chromosomes surveyed: 3 in the 5' flanking DNA (positions 130, 322, and 532 in Figure 3.2), and 1 in the second intron (1388). Each of these polymorphisms occurred on a different sequence haplotype background in a different population. Although parental material was unavailable for analysis, it is unlikely that any of these changes represent spontaneous mutations, given the estimated *de novo* mutation rate of 7.4×10^{-9} per

nucleotide per generation (Stamatoyannopoulos & Nute, 1982). Therefore, despite their low frequency in the total sample, each of these nucleotide differences probably represent true polymorphisms (i.e. changes which have persisted for more than one generation).

Fifteen (54%) of the 28 nucleotide differences observed were transitions (Table 3.1), representing a relative excess of transitions over random expectation (there are only 4 possible types of transitions (i.e. purine->purine or pyrimidine->pyrimidine), but 8 types of transversions (purine<->pyrimidine), hence 33% of random changes are potentially transitions). The observed proportion of transitions to transversions (1.15) was not, however, as high as that observed for interspecific DNA substitutions (e.g. 1.45-1.84, Li, Wu, & Luo, 1984; Blake, Hess, & Nicholson-Tuell, 1992). This lower frequency of transitions most likely derives from the absence of mutation at CG dinucleotide 'hotspots'; none of the 28 polymorphisms occurred at such sites. This discrepancy is probably due to the relative paucity of CG target sites (Poncz et al., 1983; Savatier et al., 1985) in the methylated (Perutz, 1990) ß-globin gene region, rather than to a lower than expected CG mutation rate.

Three of the 4 possible types of transitions were observed in the ß-globin region, as well as 7 of the 8 possible types of transversions (Table 3.1). T->C changes make up the largest proportion of the observed polymorphisms (32% of the total, 60% of transitions), as has been reported for pseudogene regions when CG mutations are disregarded (Blake, Hess, & Nicholson-Tuell, 1992). The most common transversion, A->T was observed in 4 out of 13 cases (31%).

Length Polymorphism

In addition to the 28 point polymorphisms, three insertions (one 1-bp, and two 2-bp) and three 1-bp deletions were observed to occur in the 5' flanking DNA region. The insertion at position 284a occurs only once in the sequences surveyed. All of the observed length polymorphisms arise in, or just adjacent to, a region of the sequence characterised by a 52-bp alternating purine-pyrimidine repeat (positions 266 to 318,

TRANSITIONS		TRANSVERSIONS	
T->C	9	A->T	4
C->T	3	T->A	1
G->A	3	T->G	3
A->G	0	G->T	0
		G->C	2
		C->G	1
		C->A	1
		A->C	1
TOTAL	15		13

Table 3.1 Summary of transitions and transversions observed in the 3 kb ß-globin gene region (5'->3' strand).

Figure A.1, Appendix A).

Previously Reported ß-Globin Sequence Polymorphism

Table 3.2 summarises the location and nature of each of the polymorphic positions identified. Other reports of these polymorphic variants are cited, including the population(s) in which each polymorphism has been described.

Thirteen of the 34 polymorphisms (38.2%) observed in Melanesia, Indonesia, and the Gambia represent new, previously undetected, polymorphic variants. Most of the newly identified polymorphisms are rare. The differences at 307, 379, 2554, 2634, 2636, 2924 and 2945, however, occur on more than one sequence haplotype (Figure 3.2) and may therefore be common enough to be observed in other populations. The failure to observe many of these changes previously may derive from the preferential analysis of the ß-globin locus and 5' regulatory regions (rather than 3' flanking DNA), motivated by the search for mutations causing haemoglobinopathies.

Seven additional nucleotide polymorphisms (-780 and -414 bp 5' of the ß-globin cap site, position 20 of the 5' untranslated region, position 30 of Exon 1, position 41 of Intron 1, position 6 of Exon 3, and position 96 of the 3' untranslated region) are also listed in Table 3.2. None of these polymorphisms were observed in the 144 sequences surveyed.

On average, 1 in 88 bp of the 3 kb human ß-globin region were found to vary polymorphically here. Although nearly 40% of these polymorphisms have not been reported previously, the observed frequency of polymorphic sites is similar to the 1 in 100 predicted on the basis of ß-globin gene cluster restriction enzyme analysis (Jeffreys, 1979). The large proportion of new variable sites detected is, most likely, a consequence of the size of the genomic region surveyed and the number of chromosomes sequenced (5.19×10^5 nucleotides were examined overall). The data do not suggest that nucleotide diversity at the ß-globin locus is higher than previously supposed.

Position in 3 kb Sequence	Position Relative to ß-globin Gene	GenBank Ref Seq	Mutated To	Comments	First Reported by	Observed in Previously	Observed in Here
68	-780	A	T		Chebloune et al., 1988	ßc and CAR ßs Genes	
130	-718	G	A				Sumatra only
138	-710	T	G		Chebloune et al., 1988	Senegal ßs gene	Senegal ßs haplotype (Gambia) only
145	-703	T	C		Semenza et al., 1984a	Albanians	Melanesia, Indonesia, and Africa
284a	-564a		AT				Gambia only
297	-551	T	C	*Rse* I RFLP	Semenza et al, 1984a, b	Europeans, SE Asians, Asian Indians, Africans	Melanesia, Indonesia, and Africa
305	-543	C	T		Chebloune et al., 1988	ßs (Benin) & ßc Genes	Gambia only
307	-541	T	C				Santo, Sumatra
318	-530	A	T		Chebloune et al., 1988	African Sickle Genes	Gambia only
319a	-529a		AT	"+ATA, -T" event	Semenza et al., 1984a	Europeans, SE Asians, Africans	Melanesia, Indonesia, Africa
320	-528	T	A		Semenza et al., 1984a	Europeans, SE Asians, Africans	Melanesia, Indonesia, Africa
322	-526	T	A				PNG Highlands Only
325	-523	T			Chebloune et al., 1988	African Sickle Genes	Gambia only
325a	-523a		T		Chebloune et al., 1988	African Sickle Genes	Gambia only
327	-521	C			Chebloune et al., 1988	Senegal ßs Gene	Melanesia, Indonesia, Africa
328	-520	T			Chebloune et al., 1988	Senegal ßs Gene	Melanesia, Africa
357	-491	A	C		Chebloune et al., 1988	Benin & CAR ßs Genes	Gambia only
379	-469	T	C				Santo, Gambia
434	-414	G	A		Chebloune et al., 1988	Central African Republic ßs Gene	
508	-340	T	C		Moschonas, de Boer, & Flavell, 1982	Europeans, Asian Indians	Melanesia, Indonesia, Africa
532	-316	T	C				Santo only
858	5' Untranslated, Nucl 20	C	T		Orkin et al., 1982	Mediterraneans, Turkish, Bulgarians	
906	Exon 1, Nucl 6	C	T	*Hgi* AI RFLP	Orkin et al., 1982; Spence et al., 1982	Europeans, SE Asians, Africans	Melanesia, Indonesia, Africa
917	Exon 1, Nucl 17	A	T		Marotta et al., 1977	Africans, Asian Indians, Arabians, Mediterraneans	Gambia only
930	Exon 1, Nucl 30	C	A	*Pst* I RFLP	Kazazian Jr. et al., 1984	Asian Indians	
994	Intron 1, Nucl 5	G	C		Kazazian Jr et al., 1984, Chang et al., 1984	Asian Indians, Chinese, Indonesians, Melanesians	Maewo, Sumatra
1030	Intron 1, Nucl 41	A	G		Kimura et al., 1983	Taiwanese (1 patient only)	
1358	Intron 2, Nucl 16	C	G	*Ave* II RFLP	Orkin et al., 1982; Spence et al., 1982	Europeans, SE Asians, Africans	Melanesia, Indonesia, Africa
1388	Intron 2, Nucl 46	T	C				Gambia only
1416	Intron 2, Nucl 74	G	T		Orkin et al 1982; Spence et al., 1982	Europeans, SE Asians, Africans	Melanesia, Indonesia, Africa
1423	Intron 2, Nucl 81	C	T		Orkin et al., 1982; Spence et al., 1982	Europeans	Island Melanesia, Africa
2008	Intron 2, Nucl 666	T	C		Orkin et al., 1982; Spence et al., 1982	Europeans, SE Asians, Africans	Melanesia, Indonesia, Africa
2198	Exon 3, Nucl 6	G	A		Kimura et al., 1983	Taiwanese (1 patient only)	
2464	3' Untranslated, Nucl 96	T	C		Indrak et al., 1992, Cai et al., 1992	Europeans, American Blacks	
2503	+50	G	A				Gambia only
2564	+101	G	C				Island Melanesia only
2634	+181	G	A				Melanesia, Indonesia
2636	+183	C	A				Melanesia, Indonesia, Africa
2792	+339	A	T	*Hin* 1I RFLP	Semenza et al., 1984a	Europeans, SE Asians, Africans	Melanesia, Indonesia, Africa
2924	+471	T	C				Gambia only
2945	+492	G	T				Melanesia, Indonesia, Africa

Table 3.2 ß-globin sequence variation – summary of polymorphisms observed and/or previously reported.

Haplotype Diversity at the Human ß-Globin Locus

The nucleotide and length polymorphisms described above were not randomly distributed among the chromosomes surveyed, but rather were found to occur in a limited number of combinations of linked sites, or sequence haplotypes, as indicated in Figure 3.2. ß-globin gene allelic diversity, as reflected in the observed sequence haplotype variation, and its relationship to the haplotype variation of surrounding restriction fragment length polymorphisms (RFLPs), is described here.

Sequence Haplotype Polymorphism

Seventy-seven sequence haplotypes were identified in the five populations surveyed (Figure 3.2). In several cases, the same sequence haplotype was observed in more than one population, so that when consolidated accordingly, the total number of unique sequence haplotypes observed was 53 (Figure 3.3). As shown in Figure 3.3, these sequence haplotypes ranged in frequency in the total sample from 22 (15.2%) to 1 (0.7%). In general, the most common haplotypes were also those which were observed in the greatest number of populations. Only one sequence haplotype, B1, was observed in all five populations. Large numbers of low frequency haplotypes were observed in Sumatra and the Gambia (Figures 3.2).

The 53 sequence haplotypes are arranged in Figure 3.3 according to overall sequence similarity and without regard to the population(s) in which the haplotypes were observed (sequences were manually aligned). As arranged, four major types of sequence haplotype, characterised by the pattern of silent nucleotide polymorphism found within the ß-globin gene (i.e. positions 906, 1358, 1416, 1423, and 2008), are apparent. Group A sequence haplotypes (numbering 11, or 20.7%, of the 53 observed) are identical to the GenBank consensus sequence at these sites, group B sequences (39.6%) differ from this haplotype pattern by a G->T change at position 1416, group C sequences (32.1%) have this difference plus, in most cases, the changes C->T at 906,

		5' Flanking																		Exon 1		Intron 1	Intron 2				3' Flanking							
		130	138	148	284a	297	306	307	318	319a	320	322	326	325a	327	328	357	378	508	532	908	917	984	1358	1388	1418	1423	2008	2583	2654	2634	2634	2792	2924	2948	FREQ
	GB Ref	G	T	T	–	T	C	T	A	–	T	T	T	–	C	T	A	T	T	T	C	A 6a	G 6h	C	T	G	C	T	G	G	G	C	A	T	G	
	1A1+2A1+3A1+4A1	.	.	C	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	14
	1A2+2A2	.	.	C	.	C	.	.	.	AT	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8
	1A5	.	.	C	.	C	.	.	.	AT	A	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
	5A9	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	2
	5A10	.	.	C	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
A	5A7	.	.	.	.	.	.	.	.	.	A	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
	5A6	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2
	1A3+2A3+4A3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6
	5A8	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2
	2A4+4A4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	2
	5A11	.	G	.	.	.	.	.	.	AT	.	.	.	.	–	–	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	2
	5B21	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	T	2
	5B22	.	.	.	.	.	T	.	.	.	A	.	.	.	.	.	C	.	.	.	.	T	.	.	.	T	.	.	.	.	.	.	.	T	1
	5B16	.	.	.	.	.	T	.	.	.	A	.	–	.	.	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	T	1
	5B17	.	.	.	AT	.	T	.	.	.	.	.	–	.	.	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	T	1
	5B18	.	.	C	.	.	T	.	.	.	.	.	–	.	.	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	T	1
	5B19	.	.	C	.	.	T	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	T	1
	5B20	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	T	1
	5B12	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	C	T	1
	5B13	.	.	C	.	.	T	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	C	T	1
	5B15	.	.	.	.	.	T	.	.	.	A	.	–	.	.	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	C	T	1
B	5B14	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	T	.	.	.	.	.	.	C	T	1
	5B11	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	T	1
	1B4+4B4	.	.	C	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	T	2
	4B9	A	.	C	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	T	1
	1B2+2B2+3B2+4B2	.	.	C	.	C	.	.	.	AT	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	T	4
	1B1+2B1+3B1+4B1+5B1	.	.	C	.	C	.	.	.	AT	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	T	22
	1B3+3B3+4B3	.	.	.	.	C	.	.	.	AT	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	T	7
	1B7	.	.	.	.	.	.	.	.	AT	A	.	.	.	–	–	.	.	C	.	.	.	.	.	.	T	.	.	.	.	.	T	.	T	1
	3B8	.	.	.	.	.	.	.	.	AT	A	.	.	.	–	–	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	T	1
	4B6+5B6	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	T	4
	1B5+4B5	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	C	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	T	3
	5C18	.	.	.	.	.	T	.	.	.	A	.	–	.	.	.	C	.	.	.	.	T	.	.	G	.	T	.	C	.	.	.	A	.	.	T	1
	4C12	.	.	C	.	C	.	.	.	AT	A	.	.	.	.	.	.	.	C	.	.	T	.	.	G	.	T	.	C	.	.	.	A	.	.	T	1
	1C2+2C2	.	.	C	.	C	.	.	.	AT	A	.	.	.	.	.	.	.	C	.	.	T	.	.	.	.	T	.	C	.	.	.	A	.	.	T	3
	3C6	.	.	.	.	.	.	.	.	AT	A	.	.	.	–	–	.	.	C	.	.	T	.	.	.	.	T	.	C	.	.	.	A	.	.	T	6
	3C7	.	.	.	.	.	.	.	.	AT	A	A	.	.	–	–	.	.	C	.	.	T	.	.	.	.	T	.	C	.	.	.	A	.	.	T	1
	4C11	.	.	.	.	.	.	.	.	AT	.	.	.	.	.	.	.	C	C	.	.	T	.	.	G	.	T	.	C	.	.	.	A	.	.	T	1
	1C4	.	.	.	.	.	.	C	.	.	A	.	.	.	.	.	.	C	C	.	.	T	.	.	G	.	T	.	C	.	.	.	A	.	.	T	1
	4C8	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	C	C	.	.	T	.	.	G	.	T	.	C	.	.	.	A	.	.	T	4
C	1C3+4C3	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	C	C	.	.	T	.	.	G	.	T	.	C	.	.	.	A	.	.	T	7
	5C19	.	.	.	.	.	.	.	.	.	A	.	.	T	.	.	.	C	C	.	.	T	.	.	G	.	T	.	C	.	.	.	A	.	.	T	1
	4C9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	C	.	.	T	.	.	G	.	T	.	C	.	.	.	A	.	.	T	3
	4C10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	C	.	.	.	.	.	G	.	T	.	C	.	.	.	A	.	.	T	1
	4C13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	T	.	.	G	.	T	.	C	.	.	A	A	.	.	T	1
	3C5	.	.	.	.	.	.	.	.	.	.	.	–	.	.	.	.	.	C	.	.	T	.	.	G	.	T	.	C	.	.	A	A	.	.	T	1
	1C1+3C1+4C1	.	.	.	.	.	.	.	.	.	A	.	.	.	–	.	.	.	C	.	.	T	.	.	G	.	T	.	C	.	.	A	A	.	.	T	5
	4C14	.	.	C	.	.	.	.	.	.	A	.	.	.	–	.	.	.	C	.	.	T	.	.	G	.	T	.	C	.	.	A	A	.	.	T	1
	4C15	.	.	.	.	.	.	.	.	.	A	.	.	.	–	.	.	.	C	.	.	T	.	C	G	.	T	.	C	.	.	A	A	.	.	T	2
	5D4	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	.	.	C	.	.	T	.	.	G	.	T	T	C	.	.	.	A	.	.	T	1
D	1D2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	G	.	T	T	C	.	C	.	A	.	.	T	1
	1D1+2D1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	T	.	.	G	.	T	T	C	.	C	.	A	.	.	T	2
	1D3	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	T	.	.	G	.	T	T	C	.	C	.	A	.	.	T	1

Figure 3.3 Condensed ß-globin sequence haplotypes, aligned to reflect sequence similarity irrespective of population of origin.

C->G at 1358, and T->C at 2008, and finally group D sequences (7.5%) are like those of group C, but with a further C->T change at position 1423. The B-, C-, and D-type sequence haplotypes divide into three, two, and two sequence subgroups, respectively, defined by haplotype similarities in the 3' flanking DNA (positions 2503 to 2945).

The frequency of sequence haplotypes of each type observed in each population, as well as in the total sample, is summarised in Table 3.3 (see also Figure 3.5). The greatest proportion of haplotypes in the combined sample were of type B (40.3%), followed by equal and smaller proportions of A- and C-type sequences (27.7%) and a much smaller number of type D sequence haplotypes (4.2%). The relative proportion of each type of haplotype in each population differed considerably: A-type sequences were most common in Island Melanesia (Santo and Maewo), whereas C-type sequences were most frequent in Sumatra, and type B sequence haplotypes were observed most often in the Highlands of Papua New Guinea (PNG) and the Gambia. In the Gambia, however, A-type sequences were the next most frequent, whereas C-type sequences were more common in PNG, so that the distribution of sequence haplotypes in these two populations was also far from identical. No D-type sequences were observed in either the PNG Highlands or Sumatra in this study.*

The non-silent nucleotide polymorphisms (positions 917 and 994) each occurred in association with two different types of sequence haplotype (Figure 3.3). The $ß^S$ (917) A ->T change was found on both a type A and a type B sequence haplotype background, while the 994 $ß^{thal}$ mutation was identified on A- and C-type sequences. The A-type $ß^{thal}$ sequence haplotype was observed in both Maewo and Sumatra. The occurrence of the same mutations on different sequence haplotype backgrounds suggests either that the same nucleotide substitutions arose independently on different sequence haplotypes, or that gene conversion (the non-reciprocal transfer of information between homologous DNA sequences) has moved each mutant from one haplotype background to another.

The distribution of silent polymorphic sites among observed sequence haplotypes

---

* Subsequent analysis of more ß-globin alleles from Sumatra has since identified several D-type sequence haplotypes in that population (J. Bond, personal communication).

Kind of Sequence Haplotype	Number Observed					
	Santo (n=48)	Maewo (n=13)	Highlands (n=24)	Sumatra (n=28)	Gambia (n=31)	TOTAL (n=144)
A	18 (37.5)	8 (61.5)	1 (4.2)	3 (10.7)	10 (32.3)	40 (27.7)
B	16 (33.3)	3 (23.1)	12 (50.0)	9 (32.1)	18 (58.1)	58 (40.3)
C	10 (20.8)	1 (7.7)	11 (45.8)	16 (57.1)	2 (6.4)	40 (27.7)
D	4 (8.3)	1 (7.7)	0 (0.0)	0 (0.0)	1 (3.2)	6 (4.2)

Table 3.3 The number of sequence haplotypes of each type observed in each population. Percent of total sample size is given in parentheses.

suggests that recombination may be occurring frequently at the locus, particularly in the region 5' of the ß-globin gene. The T->C polymorphism at position 145, for instance (Figure 3.3), was identified in all four sequence haplotype groups. As the independent recurrence of the same substitution at the same position in four sequence lineages is highly unlikely, some form of interallelic exchange (it is impossible to distinguish between reciprocal recombination or gene conversion in this case because 5' flanking polymorphisms are unavailable for comparison) must be responsible for the observed site distribution. Similar distributions, involving two or more sequence haplotype groups, are observed for all of the other common 5' flanking region polymorphisms. It is likely, therefore, that much of the 5' flanking region is affected by recombination.

There is less evidence for sequence exchanges involving silent polymorphisms in either the ß-globin gene or its 3' flanking region. Two possible instances of site-specific gene conversion, at positions 906 (haplotypes 4C10 and 1D2, Figure 3.3) and 1358 (haplotypes 1C2+2C2, 3C6, and 3C7), occur within the ß-globin gene itself. In both cases, the particular nucleotide, or allele, expected on the basis of the consensus sequence for the haplotype group (e.g. T for C- and D-type sequence haplotypes at position 906), is replaced by the other allele (e.g. C, the base observed in A- and B-type sequences); adjacent polymorphisms are, however, unaffected. The only 3' flanking polymorphism which may reflect the effects of recombination (also gene conversion) is the D-specific G->C change at 2554, which is not observed in haplotype 5D4.

Associated RFLP Haplotype Polymorphism

ß-globin gene cluster restriction enzyme haplotypes were investigated in four of the five populations surveyed for DNA sequence variation (equivalent data from Sumatra were not collected). The five RFLP sites studied, whose locations relative to the 3 kb sequenced region are shown in Figure 3.4, together make up the 3' RFLP haplotype of the human ß-globin gene cluster (Orkin & Kazazian Jr., 1984). The 12 kb RFLP haplotype includes the ß-globin gene and sites 5' and 3' of the sequenced region. 5' RFLP haplotypes (also shown in Figure 3.4), which are in linkage equilibrium with the ß-

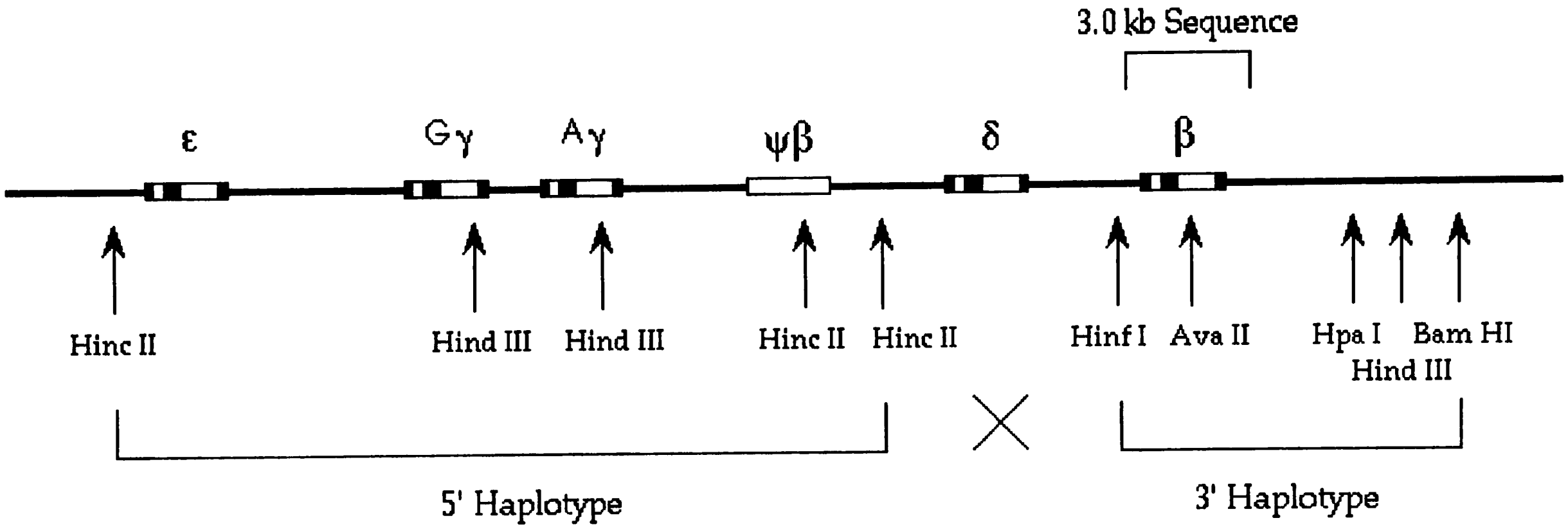


Figure 3.4 The position of the ten restriction enzyme sites which make up the 5' and 3' ß-globin gene cluster RFLP haplotypes.

globin gene (Chakravarti et al., 1984), are not reported here.

Ten 3' RFLP haplotypes were observed among the 116 chromosomes surveyed. These RFLP haplotypes, and the sequence haplotypes with which they were associated, are shown in Table 3.4. Clearly, much of the variation revealed by DNA sequencing was left undetected by restriction enzyme analysis: all but one of the 10 RFLP haplotypes were linked to two or more sequence haplotypes, and 4 (++++, +++-+, ++++-, and +-+-+) were associated with 9 or more sequence haplotypes each. In general, the most common RFLP haplotypes were linked to the largest numbers of sequence haplotypes. Sequence haplotypes linked to the same RFLP haplotype were also likely to derive from the same sequence group (i.e. A, B, C, or D), although this was not always the case. The most frequent RFLP haplotype, for instance, +++-+, was linked to A-, B-, and C-type sequences, and four other RFLP haplotypes (+++++, ++++-, +-+-+, and –+-+) were linked to two different types of sequence haplotype each. In many cases, as well, the same sequence haplotype was linked to two or more RFLP haplotypes. However, haemoglobinopathy-containing sequence haplotypes (i.e. 2A4+4A4, 5A11, and 5B22) were only linked to one RFLP haplotype each.

Combined sequence and RFLP haplotype data, rearranged to reflect the sequence haplotype alignment given in Figure 3.3, are shown in Table 3.5. As can be seen from the table, conserved RFLP haplotype patterns characterise closely related (i.e. highly similar) sequence haplotypes (those few RFLP sites which do not conform to the pattern suggested by the majority of the data are boxed). Two RFLP haplotypes, +++-+ and +++++, were found in association with most A-type sequence haplotypes. B-type sequences were predominantly linked to three RFLP haplotypes, ++–+, +++-+, and ++++-, which were each, in turn, associated with one of the three sequence subgroups identified in Figure 3.3. In contrast, both C and D-type sequence haplotypes were linked to a single RFLP haplotype, +-+-+. RFLP haplotypes were, therefore, consistent with distinctions apparent at the nucleotide level.

Differences from the expected RFLP pattern, which result either from point mutation or recombination, were observed in a number of instances (Table 3.5). The *Bam* HI site

5' *Hinf I*	SEQUENCE Haplotype	*Ava II*	*Hpa I*	3' *Hind III*	*Bam HI*	Frequency	TOTAL
	1A1+2A1					9	
	1A2					1	
	1C2+2C2					3	
	2A3					1	
	3C6					6	
+	3C7	+	+	-	+	1	28
	5A9					2	
	5B12					1	
	5B14					1	
	5B15					1	
	5B18					1	
	5B21					1	
	1B1+2B1+3B1+5B1					10	
	1B3+3B3					5	
	1B2+2B2+3B2					3	
	1B4					1	
+	1B5	+	+	+	-	1	26
	1B7					1	
	3B8					2	
	5A8					1	
	5B6					2	
	5B11					1	
	1C3					5	
	1C4					1	
	1C1+3C1					4	
	1D1+2D1					2	
+	1D2	-	+	-	+	1	17
	3C5					1	
	5C18					1	
	5C19					1	
	5D4					1	
	1A1					1	
	1A2					1	
	1A3+2A3					3	
	1B5					1	
	2A4					1	
+	5A6	+	+	+	+	2	15
	5A7					1	
	5A8					1	
	5A10					1	
	5A11					2	
	5B13					1	
	1B1+2B1+3B1					9	
-	1B3	+	+	+	-	1	12
	5B6					1	
	5B21					1	
	1A1+3A1					3	
-	1A2+2A2	+	+	-	+	5	9
	1A5					1	
	1C3					1	
-	1D1	-	+	-	+	1	3
	1D3					1	
	5B17					1	
+	5B19	+	-	-	+	1	3
	5B20					1	
-	5B16	+	-	-	+	1	2
	5B22					1	
+	2A2	+	+	-	-	1	1

Table 3.4 Linkage relationships observed between 43 sequence haplotypes and the 10 3' RFLP haplotypes. The *Ava* II site is bracketed with the name of each sequence haplotype because that site falls within the 3 0 kb sequenced region. The remaining sites lie either 5' or 3' of the sequence haplotype.

5' *Hinf I*	SEQUENCE *Haplotype*	*Ave II*	*Hpa I*	3' *Hind III*	*Bam HI*	Frequency
+	1A1+2A1	+	+	-	+	9
-	1A1+3A1	+	+	-	+	3
+	1A1	+	+	+	+	1
-	1A2+2A2	+	+	-	+	5
+	1A2	+	+	-	+	1
+	1A2	+	+	+	+	1
+	2A2	+	+	-	-	1
-	1A5	+	+	-	+	1
+	5A9	+	+	-	+	2
+	5A10	+	+	+	+	1
+	5A7	+	+	+	+	1
+	5A6	+	+	+	+	2
+	1A3+2A3	+	+	+	+	3
+	2A3	+	+	-	+	1
+	5A8	+	+	+	+	1
+	5A8	+	+	+	-	1
+	2A4	+	+	+	+	1
+	5A11	+	+	+	+	2
+	5B21	+	+	-	+	1
-	5B21	+	+	+	-	1
-	5B22	+	-	-	+	1
-	5B16	+	-	-	+	1
+	5B17	+	-	-	+	1
+	5B18	+	+	-	+	1
+	5B19	+	-	-	+	1
+	5B20	+	-	-	+	1
+	5B12	+	+	-	+	1
+	5B13	+	+	+	+	1
+	5B15	+	+	-	+	1
+	5B14	+	+	-	+	1
+	5B11	+	+	+	-	1
+	1B4	+	+	+	-	1
+	1B2+2B2+3B2	+	+	+	-	3
+	1B1+2B1+3B1+5B1	+	+	+	-	10
-	1B1+2B1+3B1	+	+	+	-	9
+	1B3+3B3	+	+	+	-	5
-	1B3	+	+	+	-	1
+	1B7	+	+	+	-	1
+	3B8	+	+	+	-	1
+	5B6	+	+	+	-	2
-	5B6	+	+	+	-	1
+	1B5	+	+	+	-	1
+	1B5	+	+	+	+	1
+	5C18	-	+	-	+	1
+	1C2+2C2	+	+	-	+	3
+	3C6	+	+	-	+	6
+	3C7	+	+	-	+	1
+	1C4	-	+	-	+	1
+	1C3	-	+	-	+	5
-	1C3	-	+	-	+	1
+	5C19	-	+	-	+	1
+	3C5	-	+	-	+	1
+	1C1+3C1	-	+	-	+	4
+	5D4	-	+	-	+	1
+	1D2	-	+	-	+	1
+	1D1+2D1	-	+	-	+	2
-	1D1	-	+	-	+	1
-	1D3	-	+	-	+	1

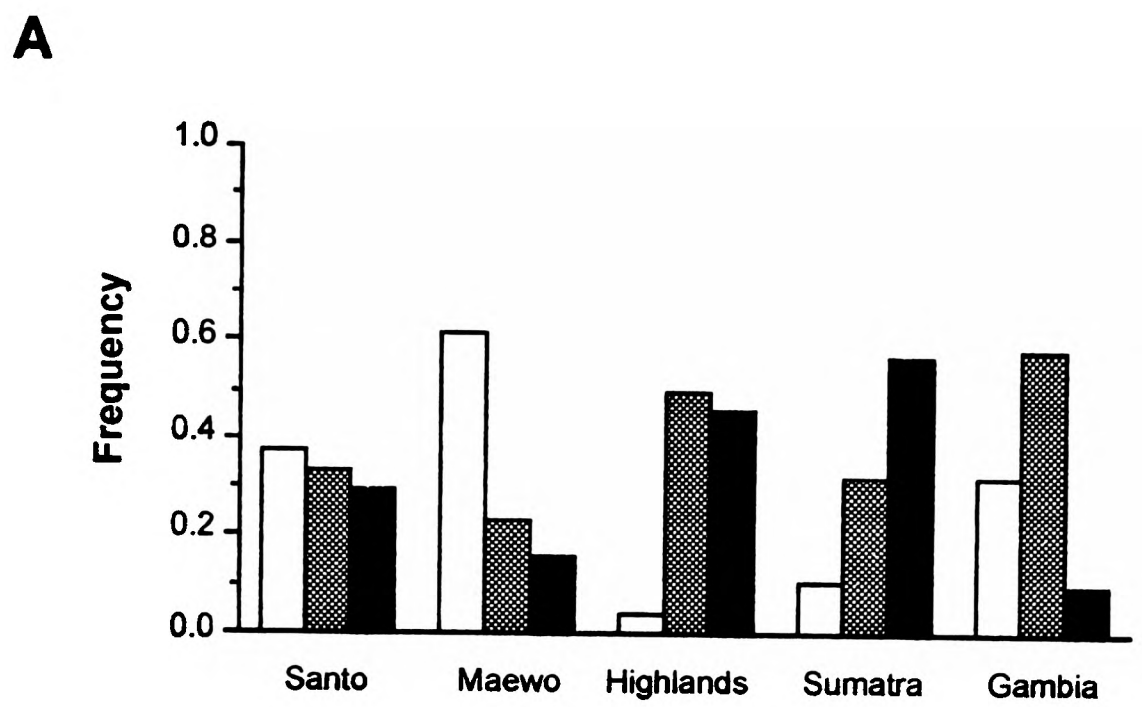
Table 3.5 Linkage relationships between sequence and RFLP haplotypes, arranged according to sequence similarity. Sites which do not conform to the majority pattern are boxed.

failed to display the expected haplotype state in 4 (3.4%) of 116 chromosomes surveyed; discrepancy rates for the *Hind* III, *Hpa* I, *Ava* II, and *Hinf* I sites were 4.3, 2.6, 8.6, and 22.4% respectively. Therefore, just as observed for the sequence haplotypes, the more 5' the RFLP site in the haplotype, the more likely it is to be involved in possible interallelic exchanges. This is particularly true of the *Hinf* I site, whose lack of strict association with adjacent sequence haplotypes is inconsistent with the majority of the data, and remarkable in light of the fact that the site lies only 175 bp 5' of the sequenced region.

Previously Reported Haplotype Variation at the ß-Globin Locus

Previous analysis, by DNA sequencing, of a small number of normal and thalassaemic ß-globin genes identified four common gene 'frameworks' at the locus (Orkin et al., 1982). These frameworks (FWs), defined by the polymorphic sites which fall within the ß-globin gene, correspond to the four sequence haplotype groups identified in Figure 3.3: thus, A-type sequence haplotypes have FW 1, B-type sequences, FW 2, C-type sequences, FW 3-Asian, and D-type sequences, FW 3, ß-globin genes. A number of studies have reported on the observed proportion of particular gene frameworks in different human populations (reviewed in Antonarakis, Kazazian Jr., & Orkin, 1985). In each population surveyed (and unlike those observed here), only three gene frameworks, or types of sequence haplotype, were identified: A, B, and D in European populations and A, B, and C in African (American Black), Southeast Asian, and Asian Indian groups. In most cases, however, gene framework typing was performed by RFLP analysis, which does not distinguish C- and D-type gene sequences, so that the presence of all four types of sequence haplotypes in these populations may have been inadvertently overlooked.

Figure 3.5 shows the observed distribution of sequence haplotypes in the populations studied here (graph A), as well as in previously investigated populations (graph B) (C and D type sequence haplotypes have been combined to make the graphs comparable). Interpopulation differences in sequence haplotype distribution identified in the current survey are also present in previously investigated samples. The available



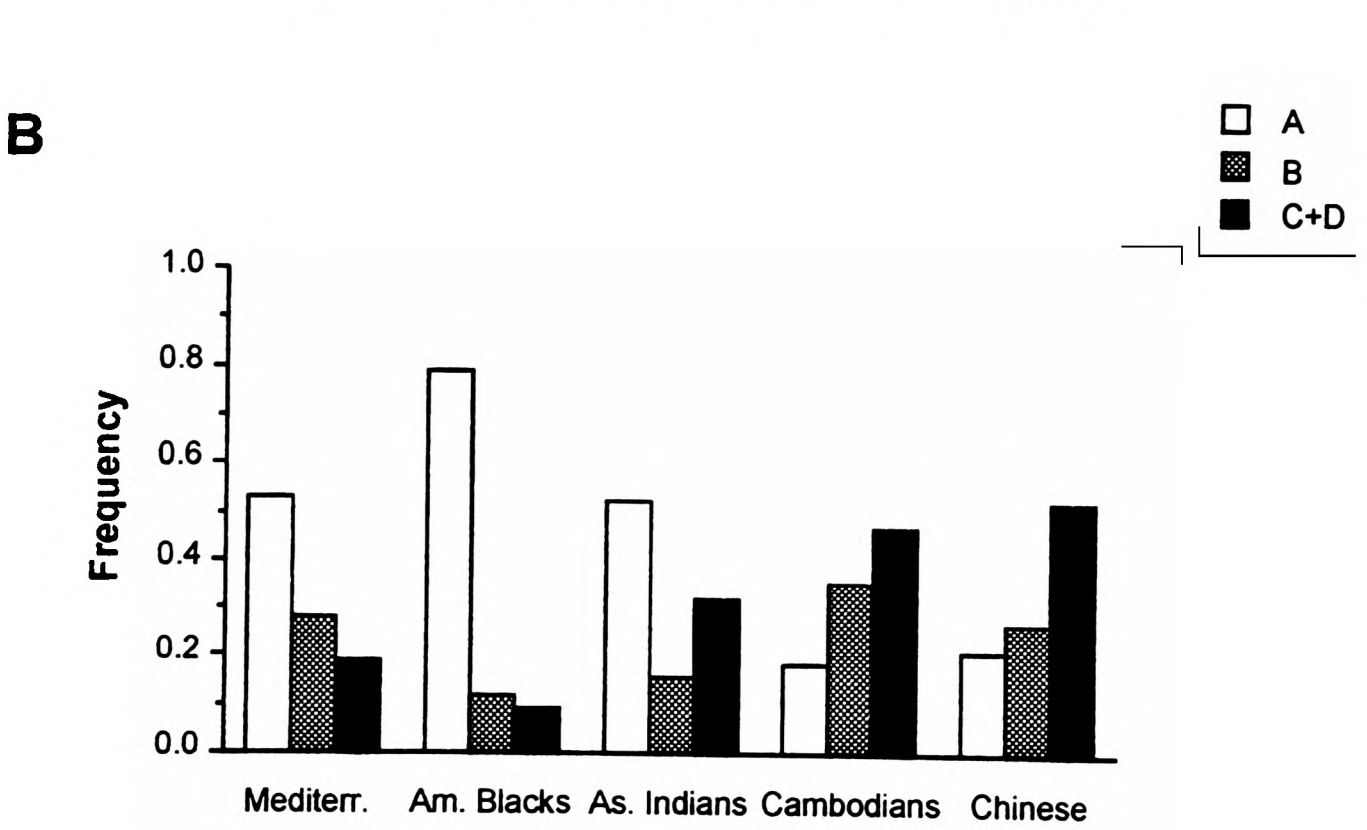


Figure 3.5 Distribution of sequence haplotype groups in various human populations. Data in B taken from Antonarakis, Kazazian Jr., & Orkin, 1985.

data suggest that closely related populations may have similar haplotype distributions. For instance, the distributions of sequence haplotypes in samples from Cambodia and China are approximately the same as that observed in Sumatra. This concordance of Asian sequence haplotype distributions could also be coincidental, however, as samples from the Mediterranean and Melanesia possess similar distributions, whereas very different distributions of sequence haplotypes are observed in the Gambia and American Blacks.

Seven different ß haemoglobinopathies ($ß^E$, $ß^S$, and five $ß^{thal}$ mutations) have been observed to occur on sequences derived from two or more sequence haplotype groups (reviewed in Matsuno et al., 1992). The $ß^S$ (position 917) mutation has been observed in association with A-, B-, and C-type sequence haplotypes (Antonarakis, et al., 1984), while the IVS 1:5 $ß^{thal}$ mutant (994) has previously been identified on both A-and C-type ß-globin sequences (Zhang et al., 1988; Lie-Injo et al., 1989). The observation of each of these mutations on two different sequence haplotypes in the sequences surveyed here is therefore consistent with the results of previous investigations.

Four 4-site 3' RFLP haplotypes (omitting the 5' *Hinf* I site) comprise 90% of normal ß-globin genes previously surveyed (Antonarakis, Kazazian Jr., & Orkin, 1985), as was observed here (Table 3.4). Conserved 3' RFLP-gene framework relationships, such as those identified for the 3 kb sequence haplotypes (Table 3.5), have also been reported previously. Two 3' RFLP haplotypes (++++ and ++-+) have been observed to be linked to A-type genes, one (+++-) to B-type genes, and another (-+-+) to both C- and D-type genes (Orkin & Kazazian Jr., 1984; Antonarakis, Kazazian Jr., & Orkin, 1985). Two of the three 3' RFLP haplotypes linked to the B-type sequences in this survey were found only in the Gambian population, which may explain why they have not been reported previously. The lack of a strict association between the *Hinf* I site and ß-globin gene sequences, a pattern observed here, has been reported by many other investigators (Kohen, Philippe, & Godet, 1982; Moschonas, de Boer, & Flavell, 1982; Antonarakis et al., 1984; Chakravarti et al., 1984; Kazazian Jr. et al., 1984); the site is often excluded from the 3' RFLP haplotype as a consequence.

Many more than the four common ß-globin alleles suggested by previous gene sequencing or RFLP analysis were identified in this study. The unexpectedly high haplotype diversity revealed by DNA sequence analysis of the 3 kb region may derive, at least in part, from randomisation of polymorphisms (via reciprocal recombination and gene conversion) in the ß-globin gene and its 5' flanking region. 5' sites (either nucleotide or RFLP) are most likely to be involved in sequence exchanges.

Comparison of Nucleotide and Sequence Haplotype Polymorphism: A Test for Recombination

As described above, DNA sequence analysis revealed the extent of nucleotide and sequence haplotype polymorphism present at the human ß-globin locus. Given the large number of sequences surveyed, it is not surprising that previously undetected variation was identified. Sequence haplotype diversity was considerably higher than that suggested by earlier studies, however, and the additional polymorphism detected appears to be due to the action of recombination within the sequenced region. The evidence for recombination, which is based solely on the observed distribution of polymorphic sites among sequence haplotypes, is suggestive but by no means definitive. Another way to test for the effects of recombination at the locus is to evaluate observed variation in terms of the expectations of the neutral mutation hypothesis (Kimura, 1968; 1983).

In the absence of recombination, the neutral theory suggests, there is a predictable relationship between observed levels of nucleotide and sequence haplotype polymorphism in a population. Kimura has shown that genetic diversity, θ, irrespective of how it is measured, is related to both the neutral mutation rate, μ, and the effective population size, N. This is because the tendency of mutation to increase diversity (which increases with increasing values of μ and N) is offset by the influence of random genetic drift (whose magnitude is inversely proportional to N), which acts to eliminate polymorphism from the population. In order to estimate θ from the observed number of polymorphic sites, the infinite sites model of mutation (Kimura, 1969), which assumes

that every mutation occurs at a previously monomorphic site, is used. The number of sequence haplotypes in a population can also be related to θ, assuming a slightly different model of mutation, called the infinite alleles model (Kimura & Crow, 1964). In this model, a new allele (or sequence haplotype) is created each time a mutation occurs. When both models of mutation hold, the number of observed polymorphic sites accurately predict the number of sequence haplotypes present in the population.

By 'shuffling' existing nucleotide polymorphisms between chromosomes, however, recombination creates new sequence haplotype variation in a manner unrelated to the rate of point mutation, so that the relationship of θ to the number of sequence haplotypes is no longer defined by the infinite alleles model. In this case, many more than the number of sequence haplotypes expected on the basis of the mutation-only estimates of θ are observed (Strobeck & Morgan, 1978). Therefore, a comparison of observed and expected numbers of sequence haplotypes (Strobeck, 1987) provides a basis for inferring the action of recombination at the locus.

Estimating Diversity from Nucleotide Polymorphism

Assuming the infinite sites model of mutation, Watterson (1975) demonstrated that the observed number of silent polymorphic sites, *s*, can be used to estimate θ. As *s* is expected to increase with increasing sample size, *n*, the estimate takes sample size into account:

$$\hat{\theta} = s \Big/ \sum_{i=1}^{n-1} 1/i \qquad (3.1)$$

For the total sample examined here, s=32 (because the non-silent polymorphisms at positions 917 and 994 are ignored) and n=144, resulting in an estimate of θ for the locus of 5.78. This observed estimate is equivalent to a level of heterozygosity per nucleotide of 0.22% ($\hat{\theta}$ divided by the 2658 effectively silent sites at the locus*). In other words, on average, 1 in 460 bp is expected to vary between any two alleles drawn

* Discussed in Chapter 4.

at random from the total sample. This level of nucleotide heterozygosity is the same as that calculated for the ß-globin gene cluster on the basis of RFLP data (0.2%, Kazazian Jr. et al., 1983), and somewhat higher than an equivalent estimate calculated for a collection of human nuclear cDNA sequences (0.08%, Li & Sadler, 1991). The significance of the latter difference is difficult to judge, however, because confidence limits for the cDNA-based estimate were not reported. Nucleotide polymorphism observed here appears consistent with previous estimates of θ at the locus and of approximately the same order of magnitude as that found in other human nuclear gene regions.

Observed and Expected Haplotype Diversity

Ewens (1972) has shown that, when the infinite alleles model is assumed, the number of sequence haplotypes, k, in a sample of size n, is the following function of θ:

$$k = 1 + \frac{\theta}{\theta + 1} + \cdots + \frac{\theta}{\theta + (n-1)} \qquad (3.2)$$

When the nucleotide-based estimate of θ, 5.78, is substituted into this equation, the expected number of sequence haplotypes at the locus is found to be 19.3 ± 3.6. The standard error of the estimate is the square root of the variance of k, given by Ewens (1972) as:

$$\mathrm{Var}(k) = k - \left[\frac{\theta^2}{\theta^2} + \frac{\theta^2}{(\theta+1)^2} + \cdots + \frac{\theta^2}{(\theta + (n-1))^2}\right] \qquad (3.3)$$

This expected number of sequence haplotypes is significantly smaller than the number of haplotypes observed in the sample of 144 chromosomes (51, when polymorphic variation at 917 and 997 is disregarded). When observed and expected numbers of sequence haplotypes are compared for three sequence subregions - the 5' flanking DNA, the gene, and the 3' flanking DNA - significant excess haplotype polymorphism is observed only in the 5' flanking DNA (Figure 3.6). Excess observed haplotype diversity and the non-uniform relationship between nucleotide and haplotype polymorphism

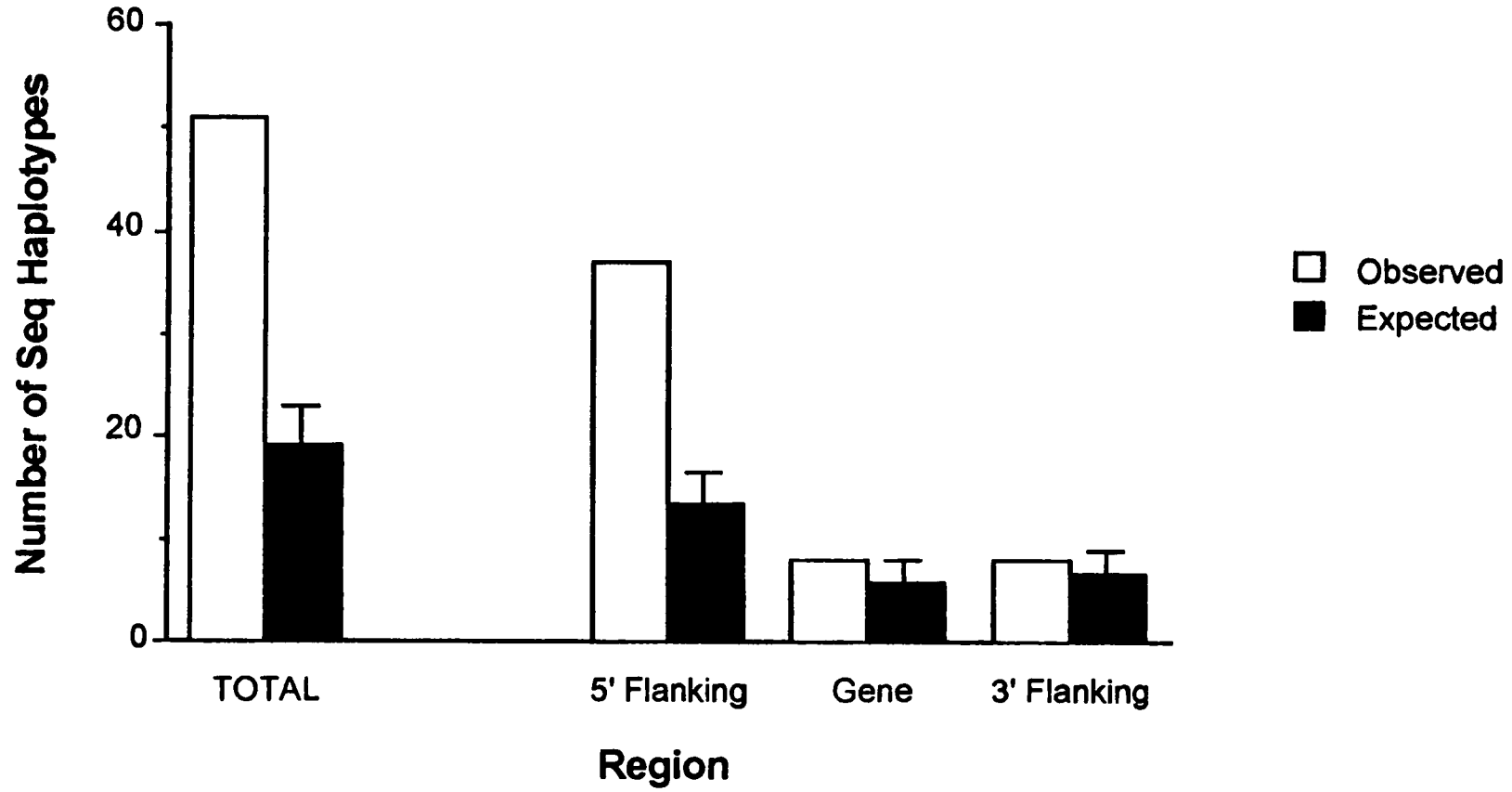


Figure 3.6 Observed and expected numbers of sequence haplotypes in the total sample (n=144).

across the sequenced region, together provide strong evidence for the impact of recombination at the locus. These results, which are consistent with inferences drawn on the basis of observed site distributions, suggest that intragenic recombination, confined predominantly to the 5' flanking region, is responsible for the unexpectedly large number of sequence haplotypes identified here.

The extent of excess haplotype polymorphism observed in the 5' flanking region varies from population to population, and when the total haplotype is considered, samples from Maewo and the Eastern Highlands of PNG do not show a significant excess number of observed sequence haplotypes (Figure 3.7). Without additional information about recombination rates or population structure, however, it is impossible to assess the significance of these differences.

Recombination at the ß-Globin Locus

This investigation of allelic sequence diversity at the human ß-globin locus has identified a remarkable cache of previously undetected nucleotide and sequence haplotype polymorphism - variation which unexpectedly suggests the extensive influence of interallelic recombination on DNA 5' of the gene. The impact of recombination identified here could only have been detected by population-based genetic study. *De novo* recombinants are rarely encountered in the course of standard molecular genetic analysis, and when they are, it is often difficult to determine precisely the nature of sequence exchanges involved. In population genetic analysis, molecular effects are implicated when similar patterns of genetic variation are observed in different geographic locales (allelic sequence relationships, unlike *de novo* recombinants, reflect the combined influence of molecular and demographic processes). Although the overall level of sequence haplotype polymorphism varied considerably among the five populations surveyed here, for example, the geographically-ubiquitous presence of excess 5' flanking region haplotype diversity was indicative of the underlying impact of recombination in that genomic region.

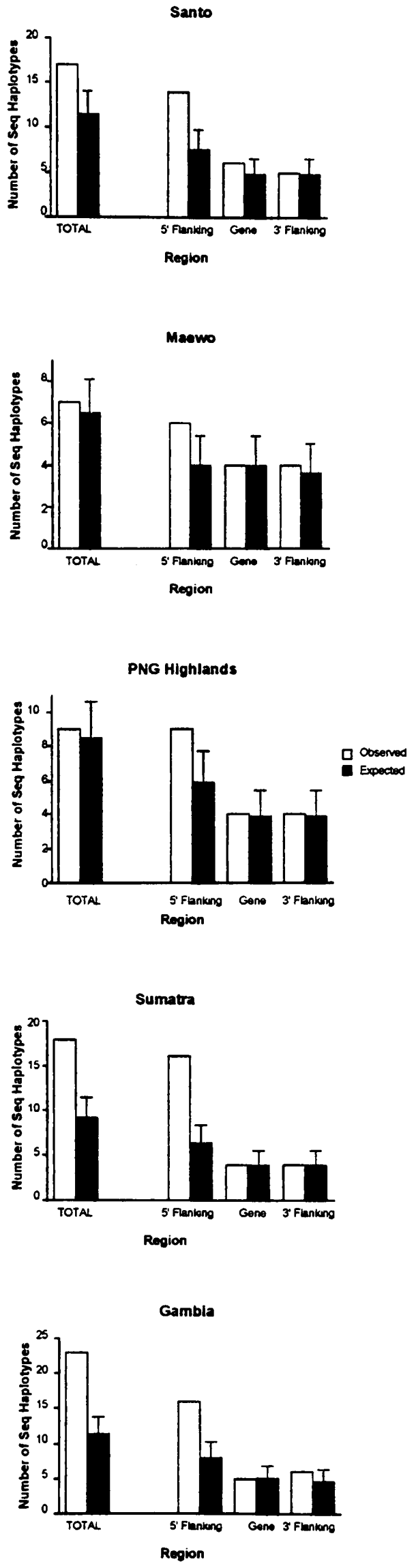


Figure 3.7 Observed and expected numbers of sequence haplotypes in each of the populations surveyed.

Watt (1972) was the first to comment on the importance of interallelic recombination (sequence exchange between alternative nucleotide sequences present at single loci) to genetic variation in populations: "as long as the minimum amount of variability is maintained at a locus," he wrote, "intragenic recombination will generate more variability by assembling all possible sequence combinations of those variations present." For all his theoretical insight, however, it is unlikely that Watt could have anticipated the peculiar pattern of interallelic exchanges apparently responsible for much of the haplotype polymorphism observed here. Excess haplotype diversity in the 3 kb region is confined almost entirely to 800 bp 5' of the ß-globin gene. The polymorphisms in this small genomic region are randomised, both with respect to one another and to the four major sequence haplotype groups, making it difficult to infer either the number or extent of sequence exchanges. Despite the inherent complexity, there is a recognisable tendency for the 5'-most polymorphisms in each population to be linked to the largest number of types of sequence haplotype, suggesting that involvement of sites in recombination may be, at least partially, a function of sequence location. These data are difficult to reconcile with the traditionally-held notion of recombination as a comparatively rare phenomenon, whose probability increases as a function of physical distance between sites.

The observed pattern of exchanges is, however, consistent with a growing body of data on recombination in yeast (and other eukaryotes), which suggest that recombination events are not evenly distributed throughout the genome (reviewed recently in Grimm, Bähler, & Kohli, 1994). Instead, largely non-recombinogenic chromosomal regions are punctuated intermittently with 'hotspots' for reciprocal exchange and gene conversion, where the majority of recombination takes place. Direct analysis of recombinants at such hotspots has shown that reciprocal recombination is frequently associated with gene conversion and that gene conversion gradients, which are assumed to reflect overlapping events that initiate at the high end of the gradient and extend for variable distances from the site of initiation (Szostak et al., 1983), are common. In most cases, such gradients extend in both directions from the hotspot, so that the likelihood of sites being involved in sequence exchanges diminishes with increasing distance from the initiation region. The length of individual gene conversion events, which determine

the extent of hotspot activity, varies considerably: Symington & Petes (1988) for instance, observed conversion tracts 4 bp to 12 kb in length at a recombination hotspot in the yeast *Saccharomyces cerevisiae*.

Although the evidence for the effects of recombination is indirect, the pattern of sequence exchanges observed here is remarkably similar to patterns associated with recombination hotspots in yeast. It is therefore not surprising to note that there is additional evidence to suggest the presence of a recombination hotspot 5' to the ß-globin gene. Antonarakis et al. (1982) were the first to identify a lack of association between RFLP polymorphisms found 5' of the ∂-globin gene and 3' of the ß-globin locus, using samples drawn from a number of human populations. He attributed the observation to an enhanced rate of recombination in a 9.1 kb segment separating the RFLP sites. Chakravarti et al. (1984) demonstrated that the observed linkage equilibrium between 5' and 3' RFLP haplotypes was consistent with rates of recombination 3 to 30 times those of flanking DNA. Subsequently, four *de novo* reciprocal recombination events (Gerhard et al., 1984; Old et al., 1986; Camaschella et al., 1988; Hall et al., 1993) were mapped to the same 9.1 kb region.

As the 800 bp 5' of the ß-globin gene falls within this 9.1 kb segment, it is conceivable that the sequence exchanges identified here mark the 3' boundary of hotspot activity. This would explain both the high frequency of inferred recombination events and the apparent tendency of the 5'-most sites to be associated with greatest number of types of sequence haplotype. Several lines of evidence suggest that the preferential site of initiation for recombination (the 'centre' of the hotspot) may lie considerably 5' of the region sequenced here. First, the 5' *Hinf* I RFLP site, which is 270 bp 5' of the first nucleotide polymorphism (position 130), is also randomised with respect to many sequence haplotypes, and (like 145) is associated with all four sequence haplotype groups. Second, previous analysis of the ß-globin 5' flanking region (introduced into *S. cerevisiae* for assay) has shown that a 1.9 kb fragment, 1400 bp 5' of the start of the sequenced region, promotes increased recombination, while a similarly sized fragment (which includes the *Hinf* I site and the 800 bp sequenced) does not (Treco, Thomas, & Arnheim, 1985).

A $(TG)_n$ repeat in the 1.9 kb region (n=16 or 17 [Poncz et al., 1983; Miesfeld, Krystal, & Arnheim, 1981]), which lies 1800 bp 5' of the 3 kb sequence, is the most likely initiation site; similar $(TG)_n$ repeats promote reciprocal recombination and gene conversion in yeast (Treco & Arnheim, 1986) and in cultured human cells (Wahls, Wallace, & Moore, 1990). Although the way in which such repeat regions promote recombination is unclear, their ability to adopt a left-handed Z-DNA conformation may be important (Blaho & Wells, 1989). The ß-globin TG repeat, however, is not absolutely required for the stimulation of sequence exchange because its deletion from the 1.9 kb region does not completely eliminate recombination (Treco, Thomas, & Arnheim, 1985). A similar lack of direct sequence dependence has been observed for other initiation sites (Schultes & Szostak, 1991; Shenkar, Shen, & Arnheim, 1991; Ponticelli & Smith, 1992) suggesting that some secondary feature, such as enhanced DNA accessibility resulting from changes in chromatin structure, may also figure prominently in a hotspot's behaviour. In this respect the location of the origin of replication for a 200 kb genomic region including the ß-globin gene cluster, which has been mapped to a 2 kb segment 1.1 kb 3' of the TG repeat (Kitsberg et al., 1993), may not be coincidental (the possible relationship between replication and recombination in yeast is discussed in Atcheson & Esposito, 1993).

If the TG repeat does represent the centre of the hotspot region, the sequence analyses presented here suggest that its effects extend well over 2.5 kb 3', into the ß-globin gene itself. The distribution of silent polymorphisms in the gene is consistent with the occurrence of at least three gene conversion events, two involving the Exon 1 transition found at a position 906 (both a C-type and a D-type sequence haplotype are affected) and one involving the transversion at 1358 (Intron 2). Therefore, although the total number of recombinants was not large enough to result in significantly more than the expected amount of haplotype diversity in the gene, sequence exchanges are having an effect on the pattern of diversity present at the locus. The distant influence of the ß-globin recombination hotspot may also explain the observed occurrence of the $ß^S$ (917) and $ß^{thal}$ (994) mutations on different sequence haplotype backgrounds; they, and 5 other ß haemoglobinopathies (four of which occur in the 5' half of the gene), appear to

have been subject to gene conversion (Matsuno et al., 1992).

It is not possible to determine with these data if the effects of the recombination hotspot extend an equal distance 5' of the (not yet identified) initiation site. However, it is likely that further sequence analysis of the ∂-ß region will go some way toward addressing this question.* These results demonstrate that a population-based survey of allelic sequence diversity can provide important information about the impact of molecular mechanisms which, like the effects of recombination identified here, may be difficult or impossible to detect directly. The way in which molecular and demographic processes interact in determining the nature of variation in particular populations is investigated in the remainder of the thesis.

---

* Such research is currently underway at the Institute of Molecular Medicine in Oxford (J. B. Clegg, personal communication).

Chapter 4

ß-Globin Sequence Polymorphism and Natural Selection

The extensive nucleotide and sequence haplotype polymorphism present at the ß-globin locus in human populations reflects the complex interaction of molecular, selective, and demographic processes. The sequence haplotype diversity described in Chapter 3, for instance, strongly suggests the influence of interallelic recombination on the genomic region 5' of the gene. Another evolutionary mechanism with potential significance for allelic sequence diversity at the human ß-globin locus is natural selection. The impact of malarial selection on the frequency and geographic distribution of non-silent ß haemoglobinopathies (mutations which modify, or impede the synthesis of, the ß chain of functional haemoglobin molecules) is well-documented. Given the presence of such haemo-globinopathies in three of the five populations investigated here, it is important to consider whether the effects of selection extend to silent polymorphism at the ß-globin locus as well.

The neutral mutation hypothesis (Kimura, 1968; 1983) makes several predictions about the effects of natural selection on DNA sequence diversity. The first, and best known, is that regions of DNA which contain sites that are important for protein expression (i.e. coding regions of genes and regulatory sequences) will vary less than non-functional genomic regions. This is because the likelihood of mutations in such regions being deleterious, therefore subject to rapid elimination by purifying selection, is proportionately greater. The remaining variation, which has no effect on gene functioning and is thus effectively neutral, is nevertheless indirectly subject to the effects of positive selection via the phenomenon of 'genetic hitchhiking' (Maynard Smith & Haigh, 1974). In other words, linkage disequilibrium between an advantageous mutation and surrounding genomic regions extends the effects of selection to nearby neutral polymorphism. As a consequence, the number, genomic distribution, and in certain circumstances, frequency of silent DNA variants, may vary according to the

nature of the selective forces involved.

There is good evidence to suggest that ß haemoglobinopathies are maintained in human populations by balancing selection (reviewed in Flint et al., 1993a; 1993b): in locales where malaria is endemic, the fitness of heterozygous carriers (who are protected from the worst effects of parasite infection but are otherwise haematologically normal) is higher than both normal and affected (i.e. homozygous) individuals. When a nucleotide substitution is subject to balancing selection, linked neutral mutations are less susceptible to loss by random genetic drift and, consequently, an excess number of silent polymorphisms are expected to accumulate in the vicinity of the selected site over time. The size of the genomic region affected depends on the degree of linkage; it is unlikely, for instance, that any increase in diversity associated with selection on sites in the ß-globin locus would extend to DNA in the gene's recombinogenic 5' flanking region. If selection is directional, rather than balancing, or the selected mutation has attained its equilibrium frequency recently, a deficit of nucleotide diversity in adjacent genomic regions is expected because unlinked variation is displaced as the selected site is driven to fixation/equilibrium. Such a selective 'sweep' (Charlesworth, 1992) will also, by changing the nature of the alleles present in the population, alter the distribution of observed site frequencies so that many more than the expected number of low frequency polymorphisms result.

In order to assess the role that natural selection plays in determining ß-globin allelic sequence diversity, the number and frequency of observed nucleotide polymorphisms among different genomic subregions of the 3 kilobase (kb) sequence were investigated in each of the four populations sampled (Espiritu Santo and Maewo were considered together as a single population, Vanuatu).

The Genomic Distribution of Silent Nucleotide Polymorphism

Heterogeneity in the observed level of nucleotide diversity (as estimated from the number of polymorphic sites) among regions of DNA may reflect variation in the level of

selective constraint, differences in the susceptibility of sites to point mutation, or the variable impact of natural selection (via genetic hitchhiking) on linked polymorphism. The contribution of selective constraint to regional differences can be disregarded if only variation at 'effectively silent' sites (Kreitman, 1983), i.e. sites which are not subject to selective constraint (or, perhaps more realistically, are constrained to the same relative degree), is considered. Any remaining heterogeneity can then be assessed assuming either a constant rate of mutation across regions, or regionally-specific mutation rates, which can be inferred from patterns of interspecific divergence.

The Level of Nucleotide Polymorphism Observed in Different Sequence Subregions

The genomic distribution of ß-globin nucleotide polymorphism is highly non-uniform (Figure 4.1): 26 (76%) of the 34 variable sites identified fall outside the ß-globin locus, in flanking regions which comprise less than 47% of the total sites surveyed, and only two (6%) of the polymorphisms found within the gene occur in coding DNA. These observations are consistent with the expectation that DNA sequences which are directly involved in protein synthesis and expression will be subject to the greatest degree of selective constraint (Kimura, 1983). In order to determine if regional differences in polymorphism exist after the effects of selective constraint are taken into account, nucleotide diversity was calculated with respect to the number of effectively silent sites in each region of interest.

As described in Chapter 3, genetic diversity, θ, can be estimated from the observed number of polymorphic sites, assuming the infinite sites model of neutral mutation (Equation 3.1). While this estimate corrects for differences in the number of sequences sampled, as formulated, $\hat{\theta}$ is still dependent on the length of the sequence surveyed (i.e. for a given sample size, more variable sites are likely to be observed in 1000 basepairs [bp] of genomic sequence than in 100), and is therefore an inappropriate measure for comparing levels of polymorphism in differently-sized regions. Regional estimates of θ can, however, be standardised by estimating diversity relative to the total number of nucleotide sites surveyed, or m:

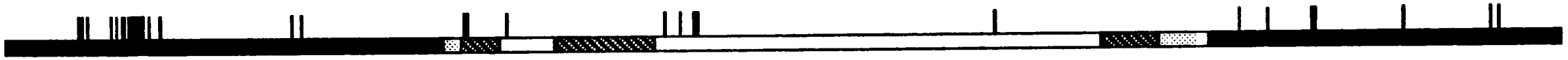


Figure 4.1 The 3 kb sequenced region, showing the location of the 34 polymorphic positions identified. Shading indicates different functional regions of the sequence, as described in Figure 3.1, Chapter 3.

$$\hat{\theta}' = \frac{\hat{\theta}}{m} \qquad (4.1)$$

The resulting estimate, which can be thought of as a measure of heterozygosity per nucleotide site, is now directly comparable to similar estimates from other regions.

In addition to standardising regional estimates of nucleotide diversity, Equation 4.1 provides a straightforward means of correcting for the effects of selective constraint because $\hat{\theta}'$ can be estimated relative to the number of effectively silent sites, rather than the total number of sites surveyed, in each genomic region. The number of such sites in the 3 kb ß-globin gene sequence was determined in accord with Kreitman & Hudson (1991), who define the number of silent sites in a region as the total number of possible silent nucleotide changes divided by three. Using this criterion, all non-coding sites, such as those found in flanking DNA and the gene's untranslated regions, all intervening (intron) nucleotides except for the conserved GT and AC splice site consensus sequences, and all four-fold synonymous sites (i.e. codon positions which can change to any other of three bases without resulting in an amino acid change) are counted as silent, as well as 33.3% of all 2-fold synonymous sites (at these sites there is only 1 possible change which will not cause an amino acid substitution, therefore they count as 1/3 of a silent site).

A total of 2658 effectively silent sites were identified in the 3000 bp sequenced. The distribution of these sites among three major subregions of the ß-globin sequence, 5' flanking, gene, and 3' flanking, is given in Table 4.1. As shown, the total number of sites surveyed and the total number of effectively silent sites is the same for regions flanking the ß-globin locus, while approximately 76% (1081) of the 1424 nucleotides found within the gene itself are effectively silent. This particular designation of silent sites is only approximate: important regulatory sequences are known to exist both 5' and 3' of the ß-globin gene (Orkin & Kazazian Jr., 1984) and certain intervening sequences (such as position 994 in Intron 1, where a non-silent ß-thalassaemia mutant occurs) are also probably subject to selective constraint. Overall, however, the number of potentially affected nucleotides form only a small proportion of the total number of

Region	Number Surveyed	Number Silent	Total Silent
5' Flanking			
5' Flanking	814	814	863
5' Untranslated	49	49	
Gene			
Exon 1	93 (60/14/19)*	24	
Exon 2	222 (141/44/37)*	52	
Exon 3	129 (81/23/25)*	33	1081
Intron 1	130	126	
Intron 2	850	846	
3' Flanking			
3' Untranslated	132	132	714
3' Flanking	582	582	
TOTAL	3000	2658	2658

*number of nondegenerate/two-fold degenerate/four-fold degenerate sites, respectively

Table 4.1 Number of effectively silent sites (Kreitman, 1983) identified in each region of the ß-globin gene sequence.

silent sites identified, so the designation is probably sufficient for the purpose of the analyses described here.

A summary of silent nucleotide polymorphism, i.e. the observed number of polymorphic sites (ignoring variation at positions 917 and 994) and resulting estimates of $\hat{\theta}'$ for the three major subregions of the ß-globin sequence, is presented for each of the surveyed populations, as well as the total sample, in Table 4.2 (because, as discussed in Chapter 5, there are no significant differences in the allelic sequence diversity observed in Espiritu Santo and Maewo, data from the two islands are considered together as Vanuatu; calculations for the Maewo portion of this combined sample are given in parentheses). The lower and upper limits of $\hat{\theta}'$ define the 95% confidence interval of the estimate, which was derived from the probability distribution given by Hudson (1990) (calculations performed by recursion, using computer programs kindly provided by R. R. Hudson):

$$P_n(s) = \sum_{i=0}^{s} P_{n-1}(s-i)Q_n(i) \qquad (4.2)$$

where

$$Q_n(s) = \left(\frac{m\hat{\theta}'}{m\hat{\theta}' + n - 1}\right)^s \frac{n-1}{m\hat{\theta}' + n - 1} \qquad (4.2a)$$

and n is number of sequences in the sample, s is the number of observed silent polymorphisms, m is the total number of effectively silent sites, and $\hat{\theta}'$ is as defined by Equation 4.1. This probability function was used, rather than the standard error of the estimate, because $\hat{\theta}'$ is not normally distributed (Kreitman & Hudson, 1991).

The overall total $\hat{\theta}'$ (for the 32 polymorphisms observed among all surveyed populations in the 3 kb sequence) is 0.00217, which corresponds to a likelihood of observing, in any one individual, 1 heterozygous site every 460 bp. The total number of polymorphic sites observed in each population (Table 4.2) was lower than that observed in the total sample (20, 16, 17, and 24 for Vanuatu, the Papua New Guinea [PNG] Highlands, Sumatra, and the Gambia, respectively). Because of differences in sample size, however, the calculated estimate of $\hat{\theta}'$ for the Gambian population was higher

Population	Sites Polymorphic	Lower	θ	Upper
Vanuatu (Maewo) (n=61(13))				
5' Flanking	10 (5)	0.00103 (0.00052)	0.00248 (0.00187)	0.00569 (0.00661)
Gene	5 (5)	0.00030 (0.00041)	0.00099 (0.00149)	0.00279 (0.00528)
3' Flanking	5 (4)	0.00045 (0.00043)	0.00150 (0.00181)	0.00422 (0.00689)
TOTAL	20 (14)	0.00080 (0.00069)	0.00161 (0.00170)	0.00322 (0.00475)
PNG Highlands (n=24)				
5' Flanking	8	0.00090	0.00248	0.00676
Gene	4	0.00024	0.00099	0.00336
3' Flanking	4	0.00037	0.00150	0.00508
TOTAL	16	0.00072	0.00161	0.00381
Sumatra (n=28)				
5' Flanking	9	0.00103	0.00268	0.00692
Gene	4	0.00024	0.00095	0.00315
3' Flanking	4	0.00036	0.00144	0.00477
TOTAL	17	0.00075	0.00164	0.00374
Gambia (n=31)				
5' Flanking	13	0.00164	0.00377	0.00886
Gene	6	0.00045	0.00139	0.00395
3' Flanking	5	0.00051	0.00175	0.00529
TOTAL	24	0.00111	0.00226	0.00483
ALL (n=144)				
5' Flanking	19	0.00205	0.00397	0.00750
Gene	6	0.00034	0.00100	0.00252
3' Flanking	7	0.00066	0.00177	0.00423
TOTAL	32	0.00122	0.00217	0.00380

Table 4.2 Summary of silent polymorphism, by regions, for each of the surveyed populations and the total sample. 'Lower' and 'Upper' refer to the lower and upper limits of the 95% confidence interval for each estimate of θ .

than that estimated for the combined sample (a graphical summary of these data is presented in Figure 4.2). The very large confidence interval associated with each estimate of $\hat{\theta}'$ (which is, in fact, typical of the measure), suggests that despite differences in the calculated estimates, total nucleotide diversity is not significantly different among populations.

When regional variation in the total sample is considered, almost equal numbers of silent nucleotide polymorphisms are observed in the gene and 3' flanking DNA (6 and 7, respectively), whereas nearly three times as many variable sites (19) are observed in the 5' flanking region (Table 4.2). The resulting estimate of $\hat{\theta}'$ for the 5' flanking DNA is four times that calculated for the gene and twice the 3' flanking region estimate. Again, however, despite considerable differences in the calculated estimates, lower and upper limits of each regional $\hat{\theta}'$ overlap. The same general pattern of regional polymorphism is observed in each population, with Gambian estimates of diversity being higher than those of other populations, particularly in the 5' flanking DNA and at the gene (Figure 4.2). Together, these data suggest that nucleotide diversity is higher in the 5' flanking region, and lower in the gene, than the overall estimate for the total 3 kb sequence. However, there are no *statistically* significant differences in the level of silent polymorphism found in particular sequence regions.

The Distribution of Nucleotide Polymorphism Among Sequence Subregions

The very large stochastic and sampling errors associated with genomic estimates of nucleotide diversity make inter-regional differences in nucleotide polymorphism difficult to detect. Another way to investigate nucleotide variability for the presence of regional heterogeneity is to test the overall distribution of polymorphic sites among genomic subregions against that expected on the basis of the neutral theory. The neutral mutation hypothesis predicts that, if the rate of mutation to silent sites is constant across genomic regions (i.e. the likelihood of any particular nucleotide mutating is the same), the number of polymorphic sites observed will be proportional to the length of the segment investigated. This is because neutral diversity is a function solely of effective population size (which does not change from one genomic region to

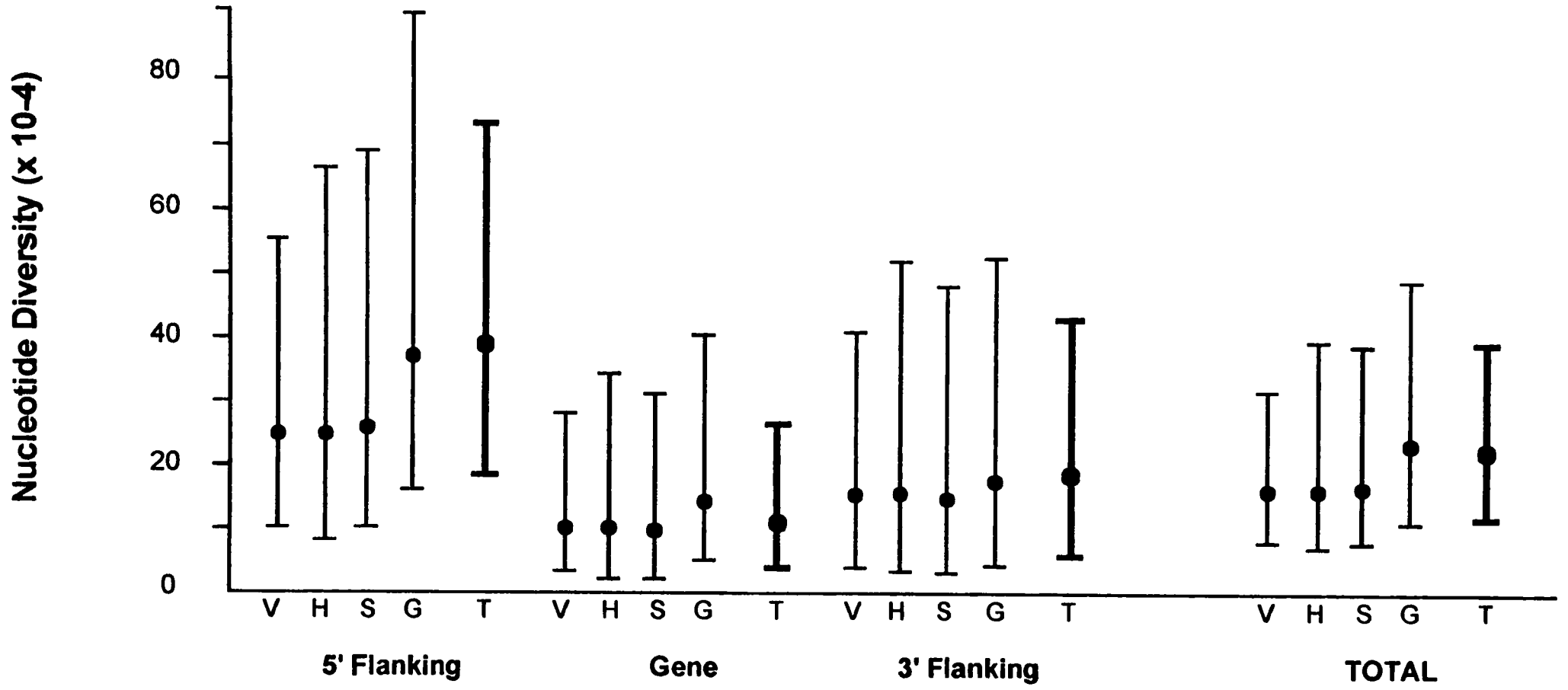


Figure 4.2 Estimates of nucleotide diversity, θ, by region, for the four populations (Santo + Maewo = V (for Vanuatu), PNG Highlands = H, Sumatra = S, and Gambia = G), as well as the total sample (T, bold brackets). The brackets represent the 95% confidence interval of each estimate.

the next in the same sample) and the rate of mutation per nucleotide per generation (Kimura, 1983). Therefore, when a constant rate of mutation is assumed, the expected distribution of nucleotide polymorphisms among subregions of the ß-globin sequence, given the total number of polymorphic sites identified, can be calculated.

Hudson (Kreitman & Hudson, 1991) has suggested a goodness-of-fit test statistic, X^2_C, for investigating the concordance of observed and expected distributions of nucleotide polymorphism in *k* regions of DNA sequence:

$$X_C^2 = \sum_{i=1}^{k} \frac{(s(i)_{obs} - s(i)_{exp})^2}{\mathrm{Var}(s(i)_{exp})} \qquad (4.3)$$

where $s(i)_{obs}$ is the observed number of polymorphic sites in the *i*th region, and $s(i)_{exp}$ is the expected number of polymorphic sites in the *i*th region, given by

$$s(i)_{exp} = \hat{\theta}'_{total} m_i \sum_{i=1}^{n-1} 1/i \qquad (4.3a)$$

Here, *n* is the number of sequences surveyed, m_i is the length in nucleotides (effectively silent sites only) of the *i*th region and $\hat{\theta}'_{total}$ is the nucleotide diversity estimate for the entire sequenced region. The variance of $s(i)_{exp}$ is calculated as

$$\mathrm{Var}(s(i)_{exp}) = \hat{\theta}'_{total} m_i \sum_{i=1}^{n-1} 1/i + (\hat{\theta}'_{total} m_i)^2 \sum_{j=1}^{n-1} 1/j^2 \qquad (4.3b)$$

When the infinite sites model of mutation (Kimura, 1969) is assumed, and there is no recombination within regions and free recombination between regions, X^2_C is approximately χ^2 distributed with *k*-1 degrees of freedom. As shown in Chapter 3, however, ß-globin sequence subregions are unlikely to recombine in this manner. If assumptions regarding recombination are violated, critical values of the test statistic are much lower than suggested by the χ^2 distribution (i.e. the test is conservative) and, as a consequence, significant differences between observed and expected distributions may remain undetected (Kreitman & Hudson, 1991). A rejection under such circumstances, of course, can be taken as strong evidence of unexpected heterogeneity in the

genomic distribution of polymorphism.

Observed and expected numbers of polymorphic sites in the 5' flanking, gene, and 3' flanking regions of the ß-globin sequence, in each population, as well as the total sample, are given in Table 4.3. Resulting estimates of X^2_C, and their respective probabilities (based upon critical values of the χ^2 distribution given in Rohlf & Sokal, 1981) are also shown. In each population surveyed, more nucleotide polymorphisms were observed in the 5' flanking region, and fewer polymorphisms in the gene, than expected (fewer than expected numbers of polymorphisms were also observed in the 3' flanking region, but to a lesser extent than that found in the gene). Not surprisingly (given the likely conservative nature of the test statistic), none of the separate population goodness-of-fit tests were statistically significant. However, the estimated probability of the Gambian X^2_C value was considerably lower than that observed in other populations, suggesting that the regional distribution of nucleotide polymorphism in the African population may be (comparatively) more heterogeneous. This difference is only a matter of degree: in the Gambia, as observed in the Southeast Asian populations, too much polymorphism is found in the 5' flanking region and too little in the gene and 3' flanking DNA.

When data from all four populations are combined, the resulting value of X^2_C *is* statistically significant (P = 0.03) due to a comparatively larger number of observed polymorphisms in the 5' flanking DNA. This finding is inconsistent with results for separate populations, and most probably reflects the consequences of analysing collectively data drawn from geographically-subdivided locales (the X^2_C test assumes that sequences are sampled from a single population of randomly mating individuals). Therefore, rather than reflecting the presence of significant heterogeneity in the genomic distribution of polymorphism (which might correspond to the influence of natural selection on the locus), the result for the total sample implies that more population-specific nucleotide polymorphisms occur in the 5' flanking region than in the gene or 3' flanking DNA.

Population	Number of Polymorphic Sites (s) Observed	Expected	$X^2 c$	P
Vanuatu (Maewo) (n=61(13))				
5' Flanking	10 (5)	6.5 (4.5)		
Gene	5 (5)	8.1 (5.7)	2.048 (0.094)	0.42 (0.95)
3' Flanking	5 (4)	5.4 (3.8)		
PNG Highlands (n=24)				
5' Flanking	8	5.2		
Gene	4	6.5	1.515	0.48
3' Flanking	4	4.3		
Sumatra (n=28)				
5' Flanking	9	5.5		
Gene	4	6.9	2.139	0.41
3' Flanking	4	4.6		
Gambia (n=31)				
5' Flanking	13	7.8		
Gene	6	9.8	3.865	0.19
3' Flanking	5	6.4		
ALL (n=144)				
5' Flanking	19	10.4		
Gene	6	13.0	7.028	0.03
3' Flanking	7	8.6		*

* = significant at the 95% level

Table 4.3 Observed and expected genomic distributions of silent nucleotide polymorphism for each of the surveyed populations, as well as the total sample.

Comparing Genomic Patterns of Polymorphism and Divergence

The available evidence fails to suggest a heterogeneous genomic distribution of nucleotide polymorphism in the populations investigated, despite the consistently greater number of variable sites observed in the 5' flanking DNA compared with the gene or 3' flanking regions. This pattern is most pronounced in the Gambian sample but not statistically significant, due perhaps to the confounding influence of recombination on the locus. Whether such differences arise as a consequence of natural selection is less clear, however. The expected distribution of polymorphic sites used to test observed nucleotide polymorphism was based on the unrealistically simple assumption of a constant rate of mutation at effectively silent sites across genomic regions. Therefore, it is possible that the observed distributions of nucleotide variation reflect genomic differences in mutation rate, rather than the extent of any particular selective effects.

A related test of nucleotide diversity, which takes account of likely differences in genomic mutation rates, has been devised by Hudson, Kreitman, & Aguadé (1987) (the HKA test). The HKA test capitalises on a very important property of the neutral mutation hypothesis, namely that both the rate of fixation of nucleotide substitutions between species and the level of polymorphism within species are functions of the underlying neutral mutation rate (Kimura, 1983). Therefore, in a region where the neutral mutation rate is high (i.e. a region in which the total proportion of mutations which are potentially deleterious is low), a large amount of interspecific sequence divergence *and* a high level of intraspecific nucleotide diversity will be observed. The HKA test uses this expected correlation to test the genomic distribution of observed nucleotide polymorphism relative to patterns of interspecific divergence.

The HKA goodness-of-fit test statistic is calculated in a similar manner to X^2_C (here, it is modified from its original presentation in Hudson, Kreitman, Aguadé, 1987, following Kreitman & Hudson, 1991, to reflect the fact that polymorphism data are available for only one of the two species investigated):

$$X^2_{HKA} = \sum_{i=1}^{k} \frac{(s_i - \hat{E}(s_i))^2}{\hat{\mathrm{Var}}(s_i)} + \sum_{i=1}^{k} \frac{(D_i - \hat{E}(D_i))^2}{\hat{\mathrm{Var}}(D_i)} \qquad (4.4)$$

where k is the number of regions being compared, s_i and D_i are the number of polymorphic and diverged sites in region i, respectively, and $\hat{E}(\ )$ and $\hat{\mathrm{Var}}(\ )$ are estimates of the expectation and variance of these sites, calculated as

$$\hat{E}(s_i) = \theta_i \sum_{j=1}^{n-1} 1/j \qquad (4.4a)$$

$$\hat{\mathrm{Var}}(s_i) = \hat{E}(s_i) + \theta_i^2 \sum_{j=1}^{n-1} 1/j^2 \qquad (4.4b)$$

$$\hat{E}(D_i) = \theta_i (T+1) \qquad (4.4c)$$

and

$$\hat{\mathrm{Var}}(D_i) = \hat{E}(D_i) + \theta_i^2 \qquad (4.4d)$$

where n is the number of sequences surveyed,

$$\theta_i = \frac{s_i + D_i}{(T+1) + \sum_{j=1}^{n-1} 1/j} \qquad (4.4e)$$

and, as derived by Shaeffer & Miller (1992),

$$T+1 = \frac{\sum_{j=1}^{n-1} 1/j \sum_{i=1}^{k} D_i}{\sum_{i=1}^{k} s_i} \qquad (4.4f)$$

X^2_{HKA} is approximately χ^2 distributed with k-1 degrees of freedom, if there is no recombination within regions and free recombination between regions. The HKA test, in common with X^2_C, is conservative if these assumptions do not hold (Hudson, Kreitman, Aguadé, 1987).

Observed and expected numbers of polymorphic and diverged nucleotide sites in the 5' flanking and gene subregions, in each population as well as the total sample, are given in Table 4.4. Estimates of interspecific divergence for the same genomic regions were based on a published comparison of human and chimpanzee (*Pan troglodytes*) ß-globin gene sequences presented by Savatier et al. (1985) (the human sequence used in the alignment was the GenBank [Bilofsky & Burks, 1988] reference sequence, a type A ß-globin gene). Unfortunately, Savatier et al. did not investigate sequence divergence in the 3' flanking region, and no other alternative chimpanzee sequence for this region was available at the time the analyses were performed. Therefore, investigation was restricted to the distribution of nucleotide polymorphism in the 5' flanking region and gene only. As observed using the X^2_C goodness-of-fit test, more than the expected number of polymorphic sites were observed in the 5' flanking region, and fewer than expected were found in the gene, in each population surveyed. The opposite pattern was observed for between-species differences. Again, however, in no population was the observed distribution of nucleotide polymorphism significantly different from that expected under neutrality, suggesting that natural selection has had no detectable effect on the pattern of nucleotide diversity at the ß-globin locus.

In contrast to the results obtained when a constant rate of mutation across genomic regions was assumed, the estimated probability of observing the distribution of nucleotide polymorphism found in the Gambia was not different from that of the other populations tested (P = 0.34, as opposed to 0.38, 0.42, 0.36 for Vanuatu, the PNG Highlands, and Sumatra respectively). The heterogeneous African distribution inferred using the X^2_C test may, therefore, derive from differences in mutation rate across genomic regions, which were only apparent in the Gambia because of that population's higher overall level of nucleotide diversity. In common with the X^2_C goodness-of-fit test results, analysis of the total data set using the X^2_{HKA} test statistic did suggest a nearly significant difference (P = 0.07) between observed and expected distributions of nucleotide polymorphism, with more than the expected number of nucleotide polymorphisms occurring in the 5' flanking region, and fewer in the gene (Table 4.4). Again, however, this finding is not likely to reflect the effects of natural selection. Instead, the observed patterns are probably a result of the combined assessment of

Population	Within Species (s)		Between Species (D)		X^2_{HKA}	P
	Observed	Expected	Observed	Expected		
Vanuatu (Maewo) (n=61(13))						
5' Flanking	10 (5)	8 (4.8)	14 (14)	16 (14.3)	1.147 (0.021)	0.38 (0.90)
Gene	5 (5)	7 (5.3)	16 (16)	14 (15.8)		
PNG Highlands (n=24)						
5' Flanking	8	6.3	14	15.7	0.917	0.42
Gene	4	5.7	16	14.3		
Sumatra (n=28)						
5' Flanking	9	7.0	14	16.0	1.242	0.36
Gene	4	6.0	16	14.0		
Gambia (n=31)						
5' Flanking	13	10.5	14	16.5	1.329	0.34
Gene	6	8.5	16	13.5		
ALL (n=144)						
5' Flanking	19	15.0	14	18.0	3.317	0.07
Gene	6	10.0	16	12.0		

Table 4.4 Results of the Hudson, Kreitman, Aguadé (1987) test of neutrality for each of the surveyed populations, and the total sample.

variation drawn from geographically-subdivided, and genetically-differentiated, populations (Chapter 5).

Although unlikely to derive from the effects of natural selection, the genomic distribution of nucleotide polymorphism in the total sample is indicative of underlying genetic differences among ß-globin subregions. To explore further the relationship between polymorphism and divergence in the total sample, the 25 nucleotide polymorphisms identified in the gene and 5' flanking DNA were partitioned into four arbitrary subregions containing roughly equal numbers of variable sites (Table 4.5, Comparison 1). The boundaries of these subregions, which separate genic sequences from three regions of 5' flanking DNA, are shown in Figure 4.3. Large differences in the size of these regions, which otherwise contain equal numbers of silent polymorphic sites, immediately suggest heterogeneity in the distribution of polymorphism. When variation in these regions is compared to that expected on the basis of observed sequence divergence, significantly greater numbers of polymorphisms are observed in two of the three 5' flanking regions (5'-1 and 5'-2) than in the gene (P = 0.04 and 0.06, respectively). This result suggests that more polymorphisms arise 500 bp 5' of the gene than elsewhere in the 3 kb sequence. A 52-bp alternating purine-pyrimidine repeat, which may be susceptible to both length and point mutation, is found in the midst of the affected subregions. It is unlikely that a higher underlying mutation rate is the sole explanation for the pattern of population-specific polymorphism identified, however, because a correspondingly high level of interspecific divergence in these regions is not also observed.

In summary, there is no convincing evidence to suggest that natural selection has had an impact on the distribution of nucleotide polymorphism at the human ß-globin locus. Estimates of nucleotide diversity in 3 subregions of the ß-globin sequence, calculated from the observed number of polymorphic sites, are not significantly different. In addition, the distribution of nucleotide polymorphisms among sequence subregions is not unexpectedly heterogeneous, whether or not differences in regional mutation rates are taken into account. Despite these findings, estimates of nucleotide diversity are consistently higher in the 5' flanking region, and more than the expected number of

Comparison	Within Species (s)		Between Species (D)		X^2_{HKA}	P
	Observed	Expected	Observed	Expected		
1						
5' -1	6	4.1	3	4.9		
5' -2	7	4.5	3	5.5	5.573	0.17
5' -3	6	6.4	8	7.6		
Gene	6	10.0	16	12.0		
2						
5'-1	6	3.5	3	5.5	3.515	0.06
Gene	6	8.5	16	13.5		
3						
5'-2	7	4.1	3	5.9	4.284	0.04 *
Gene	6	8.9	16	13.1		
4						
5'-3	6	4.7	8	9.3	0.765	0.45
Gene	6	7.3	16	14.7		

* = significant at the 95% level

Table 4.5 Results of the HKA test for sub-regions of the 5' flanking DNA (using polymorphic sites observed in the total sample, n=144).

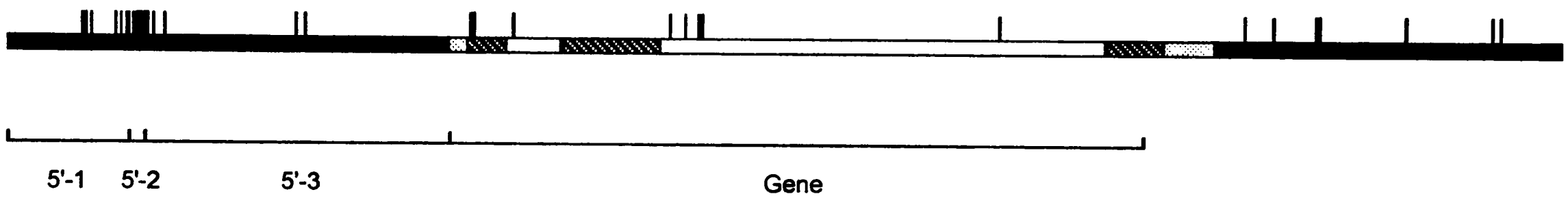


Figure 4.3 The boundaries of the four subregions compared using the HKA test.

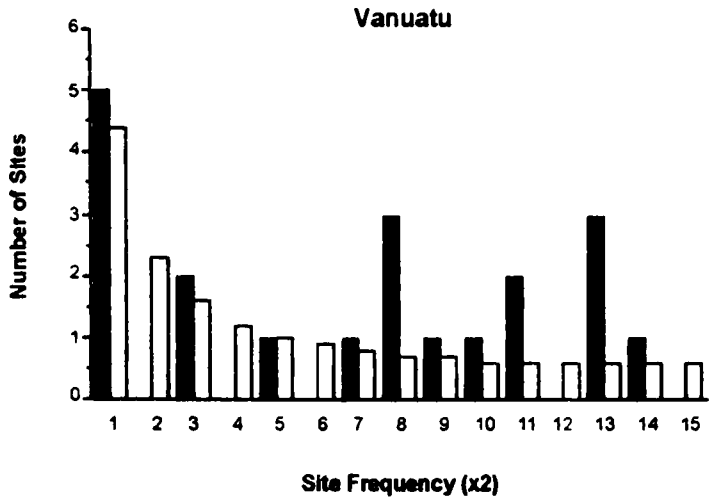
polymorphisms are observed in the 5' flanking DNA, in all populations.

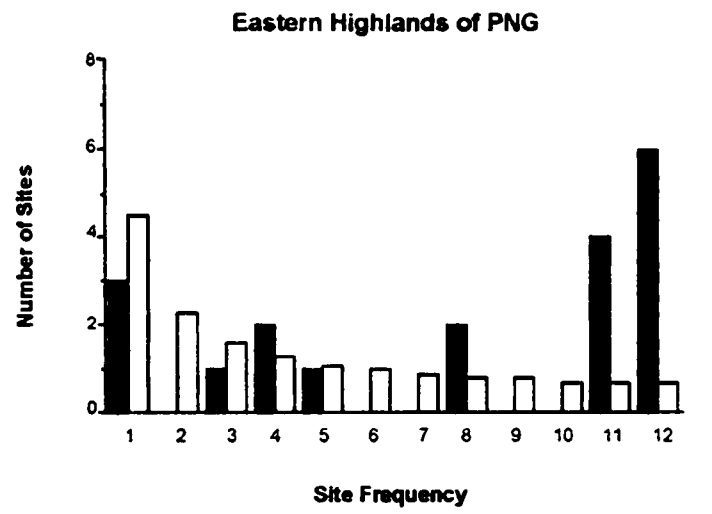
The Frequency of Observed Polymorphic Sites

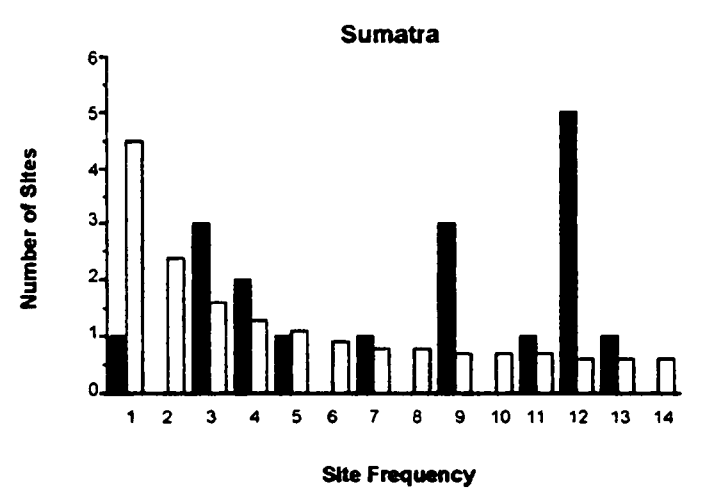
There is little evidence to suggest the influence of natural selection on the genomic distribution of nucleotide polymorphism at the human ß-globin locus. Given the presumed recent impact of malaria in human populations (Flint et al., 1993a), it is possible that selection has not been acting on the locus long enough to affect the number of sites found in particular genomic regions. By altering rapidly the frequency of a selected allele in a population, however, natural selection can modify the frequency of nucleotide variants in such a way that the effects of selection are apparent soon after on-set. When a selectively-advantageous mutation arises, the frequency of the mutant (and any linked neutral polymorphism) in the population will increase rapidly, often in only a small number of generations. The end result is a few sites of high to intermediate frequency (those that were linked to the selected mutation) and a proportionately greater number of low frequency polymorphisms, a pattern which erodes once the selected mutant ceases to be advantageous, or is fixed in the population. The recent impact of natural selection may therefore be inferred by comparing observed and expected frequencies of nucleotide polymorphisms in a population.

Observed and Expected Site Frequency Distributions

The observed frequency of silent polymorphic sites in each population surveyed is summarised in Figure 4.4 , where the number of polymorphisms which occur *i* times in a sample (*i* ranging from 1 to *n*/2, where *n* is the total number of sequences surveyed) are shown (dark bars of the histograms). In plotting these distributions, each nucleotide present at a variable position was counted as a separate site: for example, if a G->C transversion occurred five times in a sample of 15 sequences, the C allele was counted as one site of frequency 5, and the G allele was counted as another site of frequency 10. Because every nucleotide polymorphism identified here was dimorphic, the resulting distribution of frequencies was symmetrical, and therefore only the first half of







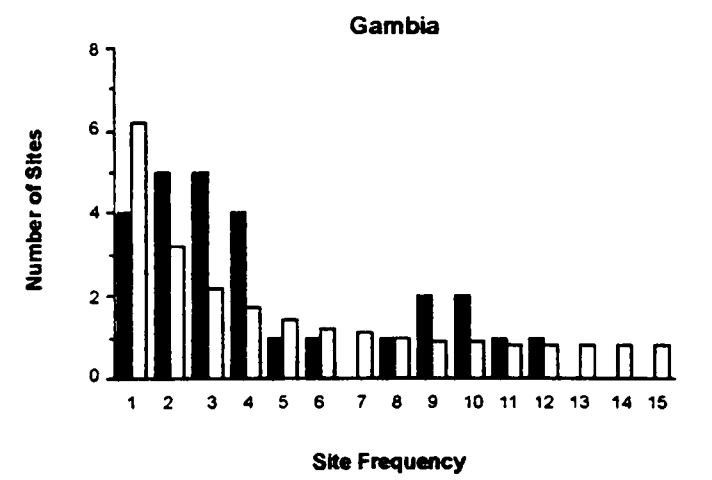


Figure 4.4 Observed (dark bars) versus expected (light bars) frequencies of silent polymorphisms for each of the populations surveyed (Espiritu Santo and Maewo combined in Vanuatu). The Vanuatu distribution is compressed (see text) for comparative purposes.

each distribution is presented. In addition, the distribution of site frequencies in Vanuatu is compressed by two (i.e. sites with frequencies 1 and 2 in the sample of 61 are shown in column 1, 3 and 4 in 2, and so on) to make the histogram more directly comparable to those of the other populations. As Figure 4.4 illustrates, each sample has a unique silent site frequency distribution. In the PNG Highlands and Sumatra, more intermediate frequency than low frequency sites were observed, whereas in the Gambia, the majority of observed polymorphisms occur at a low frequency. These population-specific site distributions suggest that different evolutionary forces (including possibly, natural selection) are influencing the nature of variation in each sample.

The expected distribution of sites in each population, assuming the infinite sites model of neutral mutation (Kimura, 1969), is also shown in Figure 4.4 (light bars of the histograms). Expected site frequencies were estimated using the total calculated diversity of the ß-globin gene region, derived from the observed number of polymorphic sites in each sample, as suggested by Tajima (1989a):

$$G_n(i) = \hat{\theta}'_{\text{total}} m\left(\frac{1}{i} + \frac{1}{n-i}\right) \qquad (4.5)$$

Here, $G_n(i)$ is the expected number of sites whose frequency is i/n in a sample of n DNA sequences, $\hat{\theta}'_{\text{total}}$ is the estimate of $\hat{\theta}'$ for the entire 3 kb sequence (Table 4.2), and m is the total number of effectively silent sites surveyed (i.e. 2658 for the 3 kb sequence). Clearly, none of the observed site frequency distributions adhere strictly to the pattern expected under neutrality. In order to evaluate the significance of apparent differences between observed and expected site frequency distributions, the Tajima test of neutrality (1989a) was used.

The Tajima Test of Neutral Mutation

As discussed elsewhere in the thesis, the genetic diversity of a DNA region may be estimated either from the observed number of polymorphic sites in the sample ($\hat{\theta}$, Equation 3.1) or the average number of pairwise differences between sequences ($\hat{\pi}_i$,

Equation 5.3). The first estimate assumes an infinite sites distribution of nucleotide polymorphism (like that depicted in Figure 4.4), whereas the second takes the actual frequency of polymorphic sites in the sample into account. Therefore, if the observed site frequency distribution is similar to that expected under the infinite sites model (i.e. the distribution of alleles in the population is neutral), the two estimates of diversity will be equal (as long as the sample is drawn from a randomly mating population). This expected relationship forms the basis of a test of the neutral mutation hypothesis proposed by Tajima (1989a). The test statistic, *D*, is calculated from the difference between the estimates of diversity in the following manner:

$$D = \frac{d}{\sqrt{\hat{V}(d)}} = \frac{\hat{\pi}_i - \hat{\theta}}{\sqrt{e_1 s + e_2 s(s-1)}} \qquad (4.6)$$

where $\hat{\pi}_i$ and $\hat{\theta}$ are as previously defined (Equations 5.3 and 3.1), s is the total number of silent polymorphic sites, and e_1 and e_2 are functions of the sample size, n (derived from a series of equations given in Tajima (1989a)):

$$e_1 = \frac{n+1}{3(n-1)x} - \frac{1}{x^2} \qquad (4.6a)$$

$$e_2 = \frac{2(n^2+n+3)x^2 - 9(n^2+n-2)x + 9(n^2-n)y}{9(n^2-n)x^2(x^2+y)} \qquad (4.6b)$$

where

$$x = \sum_{i=1}^{n-1} 1/i \quad \text{and} \quad y = \sum_{i=1}^{n-1} 1/i^2$$

If the calculated value of *D* is significantly different from zero (as determined by the confidence limits of the test statistic provided by Tajima in Table 2 of his paper),* the observed distribution of site frequencies is inconsistent with that predicted by the neutral theory. A significantly negative *D* suggests that more low frequency sites than expected are present in the sample, whereas a significantly positive value of *D* suggests that fewer

---

* Tajima suggests that his test will be conservative if the locus under consideration is subject to the effects of recombination. As a consequence, when the test is applied to nuclear DNA loci (such as the ß-globin gene), probability values less than 0.1 are regarded as statistically significant.

than the expected number of low frequency sites are observed.

Estimates of $\hat{\pi}$, and $\hat{\theta}$, and the resulting value of *D*, for the total sequence as well as the 5' flanking, gene, and 3' flanking subregions, in each population and the combined sample, are given in Table 4.6. As shown, the only population in which the observed frequency of polymorphic sites was significantly different from neutrality was Sumatra (*D* = 1.75 for the total sequence, $P < 0.1$); this result derives from a significantly positive value of *D* for the gene. Nearly all of the estimates of *D* (for different populations and subregions) were positive, suggesting that overall, fewer low frequency sites are present at the ß-globin locus in the populations surveyed (this finding is consistent with the site frequency distributions presented in Figure 4.4). A relative deficit of low frequency sites can arise as a consequence of balancing selection: silent polymorphisms which are linked to a selectively-maintained mutation are held in the population at a higher than expected frequency while remaining unlinked variability is distributed normally - the net effect is more than the expected number of intermediate frequency variants. This effect is not expected to extend into regions subject to the effects of interallelic recombination.

ß haemoglobinopathies, which may be presumed to be subject to the effects of natural selection, are present in three of the five populations surveyed: Maewo ($ß^{thal}$), Sumatra ($ß^{thal}$ and $ß^{E}$), and the Gambia ($ß^{S}$). As the impact of selection on variation in these populations is expected to be similar (each is believed to infer to carriers important protection from the worst effects of malarial infection [Flint et al., 1993a; 1993b]), it is useful to compare Tajima test results for these populations. Overall estimates of *D* were considerably different (Table 4.6): in Maewo and the Gambia, *D* was close to 0, suggesting little or no difference in the observed and expected frequencies of sites in these populations, whereas the Sumatran *D* was one of the largest values calculated. When variation in the gene only was considered, estimates for Maewo and the Gambia were both negative (the sole negative values of *D* in the current study) while the *D* calculated for Sumatra was significantly positive. Negative values of *D* are consistent with the recent increase in frequency of a selected mutation, which has certainly occurred in Maewo (Hill et al., 1988), and could possibly have occurred in the Gambia.

Population	Region			TOTAL
	5' Flanking	Gene	3' Flanking	(All Regions)
Vanuatu (Maewo) (n=61(13))				
π	2.75 (2.28)	1.76 (1.38)	1.51 (1.33)	6.03 (5.00)
θ	2.14 (1.61)	1.07 (1.61)	1.07 (1.29)	4.27 (4.51)
D	0.79 (1.48)	1.50 (-0.50)	0.96 (0.12)	1.25 (0.45)
PNG Highlands (n=24)				
π	3.14	1.41	1.41	5.97
θ	2.14	1.07	1.07	4.28
D	1.50	0.86	0.87	1.40
Sumatra (n=28)				
π	3.38	1.73	1.46	6.58
θ	2.31	1.03	1.03	4.37
D	1.46	1.79 *	1.11	1.75 *
Gambia (n=31)				
π	3.83	1.12	1.31	6.26
θ	3.25	1.50	1.25	6.01
D	0.58	-0.71	0.13	0.15
ALL (n=144)				
π	3.63	1.74	1.56	6.93
θ	3.43	1.08	1.26	5.77
D	0.16	1.29	0.52	0.59

* = P<0.1. All other *D* values are P>0.1.

Table 4.6 Results of the Tajima test of selective neutrality (Tajima, 1989a), by regions, for each of the surveyed populations and the total sample.

The finding of a much larger *D* in Sumatra is inconsistent with the other results, and may reflect the fact that variation in the population is affected by factors other than natural selection.

One possible alternative explanation for the Sumatra findings is population admixture. A large positive *D* value was observed for the gene when variation in the total sample (which can be thought of as 'effectively' admixed, by virtue of its being comprised of alleles drawn from several genetically-distinct populations) is considered. In Maewo, the PNG Highlands, and the Gambia, however, estimated *D* values were greatest in the 5' flanking region, a pattern which may in some way be associated with the higher rate of recombination in that genomic region. There is, therefore, a tentative indication from the available data that Vanuatu and Sumatra represent admixed populations. Different *D* values in these populations might then reflect the recency or extent of admixture in each locale. This hypothesis, while attractive, is likely to remain unproven until the Tajima test is applied to the investigation of ß-globin sequence diversity in other populations.

In summary, while there is some evidence to suggest that natural selection has affected the frequency of silent polymorphisms present at the locus, inconsistencies in the results derived for the 3 populations most likely to be subject to the effects of malarial selection (Maewo, Sumatra, and the Gambia) suggest that demographic, rather than selective, phenomena have had a predominant influence on the observed site frequency distributions. Overall, more than the expected number of intermediate frequency sites were observed in all surveyed populations. Positive *D* values did not derive from the impact of balancing selection on the gene, however; instead these findings most often reflected the greater than expected number of intermediate frequency sites in the 5' flanking DNA, the genomic region least likely to be influenced by genetic hitchhiking of selected sites in the ß-globin locus. The only significant difference from neutrality was observed in Sumatra, but comparison with data from other populations suggested that this result was more likely to be due to the effects of population admixture than selection. In Maewo and the Gambia, the presence of selectively-favoured ß haemoglobinopathies appears to have led to fewer, rather than

more, than the expected number of intermediate frequency sites in the gene.

Natural Selection and Recombination at the ß-Globin Locus

There is no doubt that natural selection has influenced the nature of non-silent variation at the ß-globin locus in three of the five populations surveyed: ß haemoglobinopathies ($ß^{thal}$, $ß^{E}$, and $ß^{S}$), which are likely to confer to carriers protection against malarial infection (Flint et al., 1993a; 1993b), are observed at frequencies of 6% or more in Maewo, Sumatra, and the Gambia. Somewhat unexpectedly, however, polymorphism in only one population, Sumatra, was significantly different from neutral theory expectations - no other evidence for the effects of selection on either the genomic distribution of polymorphic sites, or the observed distribution of polymorphisms among sequence haplotypes, was observed. These findings suggest either that allelic sequence diversity at the ß-globin gene is unaffected by the presence or absence of selected mutations in particular populations (i.e. is not subject to the effects of genetic hitchhiking) or that the available tests are not powerful enough to distinguish the effects of selection from those of other population-specific phenomena.

There are good reasons to suspect the efficacy of at least one of the approaches employed here to test for the effects of natural selection. At genes such as the alcohol dehydrogenase locus in *Drosophila melanogaster*, where the impact of selection can be inferred from the temporal and spatial distribution of electrophoretic polymorphism (van Delden, 1982) as well as from underlying patterns of DNA sequence variation (Kreitman & Aguadé, 1986; Hudson, Kreitman, & Aguadé, 1987; Kreitman & Hudson, 1991), the Tajima test of selective neutrality has consistently failed to detect any significant departures from neutral equilibrium (see also Martin-Campos et al., 1992; Langley et al., 1993; Begun & Aquadro, 1993; Begun & Aquadro, 1994). In addition, when significant Tajima test findings *have* been reported, for example the significantly negative *D* values estimated on the basis of restriction fragment length polymorphism (RFLP) haplotype variation at the human mitochondrial DNA (mtDNA) locus (Merriwether et al., 1991), they, like the Sumatran results identified here, appear to be

better explained by demographic rather than selective phenomena (in the case of mtDNA, there is evidence to suggest the influence of recent population expansion, instead of population admixture, on observed genetic diversity).

The HKA test, while more successful than the Tajima test in identifying the effects of both balancing and directional selection on variation at *Drosophila* loci, has nevertheless failed to reveal the impact of malarial selection on polymorphism at the human ß-globin gene. This failure derives not from any inherent limitation of the test, but rather from the comparatively recent influence of malaria in human populations. It is generally believed that *falciparum* malaria has only been an important selective force, in terms of its effects on human morbidity and mortality, in the last 2,000 to 5,000 years (Livingstone, 1989; Flint et al., 1993b). Even if we assume that each of the advantageous mutations identified here was present in its respective geographic locale when malarial selection began, this amount of time is simply not long enough to have resulted in a detectable increase in neutral polymorphism in linked genomic regions: given the estimated rate of sequence divergence at effectively silent sites of 0.3% per million years,* less than one new silent polymorphism would be expected to arise in the 3 kb region in a 5000 year period. The relatively low equilibrium frequency of observed haemoglobinopathies also makes it unlikely that any appreciable reduction in the level of unlinked sequence diversity at the ß-globin locus accompanied the establishment of the mutations in affected populations.

Significant perturbation of the genomic distribution of silent nucleotide polymorphism would clearly only result if malaria had maintained selected ß haemoglobinopathies in human populations for considerably longer periods. Of course, subsequent introduction of more advantageous mutants, which would rapidly supplant older variants (and therefore any associated excess polymorphism), might render even the long-standing impact of balancing selection undetectable. Fitter genotypes can arise as a consequence of *de novo* mutation, whose effects are random and unpredictable, or

---

* This rate of sequence divergence is based on the number of sequence differences observed between human and chimpanzee ß-globin genes (Savatier et al., 1985; J. Bond, personal communication), assuming a 5 million year old nearest common ancestor (Saitou & Ueda, 1994).

may be 'constructed' by the transfer of existing mutations to new chromosomal backgrounds via interallelic recombination. As recent experimental work involving the *in vitro* simulation of recombination in *Escherichia coli* has demonstrated (Stemmer, 1994), the latter tactic can quickly generate alleles with increased selective fitness, owing largely to the fact that advantageous mutations are likely to confer the same or higher level of protection in a different sequence context (Smith, 1994). Thus, at a locus as susceptible to the effects of selection as ß-globin, the influence of interallelic recombination could be, adaptively, extremely important.

In fact, there is evidence to suggest that recombination (in particular, gene conversion), in conjunction with malarial selection, *has* significantly influenced the distribution of ß haemoglobinopathies in human populations. For example, the HbS mutation (the A->T transversion observed at position 917 in the Gambia), is observed to occur predominantly on 5 different sequence haplotype backgrounds, 4 found in West Africa (Pagnier et al., 1984; Lapouméroulie et al., 1992), and a fifth confined to Saudi Arabian and Asian Indian populations (Kulozik et al., 1986). The regional specificity of these 5 types of sickle chromosomes, which differ from one another by more than a single crossing-over event, has been used to argue that the HbS mutation arose independently 5 times (Pagnier et al., 1984; Cheboune et al., 1988; Trabuchet et al., 1991; Öner et al., 1992). However, as Flint et al. (1993a; 1993b) have pointed out, neither the haplotype variation associated with each regional mutation, nor the geographical distribution of the different types of sickle genes, is consistent with the multiple origin hypothesis. Rather, the observed HbS pattern is more likely to have resulted from the conversion of the mutation onto distinct haplotype backgrounds, a possibility first suggested by Antonarakis, et al. (1984).

The repeated conversion of a small region of DNA in the first exon of the ß-globin gene is plausible, both in terms of similar observations involving other haemoglobinopathies (Matsuno et al., 1992) as well as the apparent impact of interallelic recombination on silent polymorphism at the locus (Chapter 3). Although gene conversion per se does not explain the remarkable geographic specificity of each major type of African sickle chromosome, the differential fitness of the resulting recombinant haplotypes might. It is

well known, for instance, that the severity of sickle cell disease can differ according to the type of sickle chromosomes a patient inherits, due in part to the co-inheritance of linked genes controlling fetal haemoglobin expression* (Labie & Nagel, 1987). Without medical intervention, individuals inheriting two Central African Republic (CAR) $ß^S$ chromosomes rarely live to adulthood, whereas patients inheriting Senegalese or Saudi Arabian sickle genes are typically far less ill; the other major type of HbS chromosome, Benin, is intermediate in severity in homozygotes (Powars, 1991). In this context, the observation of two types of sickle chromosome in the Gambia is consistent with the recent increase in frequency of the less deleterious Senegal allele following conversion from an introduced Benin HbS haplotype. The genetic composition of other populations in which HbS is found will need to be investigated before the nature of events surrounding the origin and geographic dispersal of the sickle cell mutation in human populations can be deciphered.

The highest levels of nucleotide diversity in each population surveyed were observed in the recombinogenic 5' flanking DNA, suggesting that interallelic recombination also affects silent polymorphism at the ß-globin locus. The influence of recombination is most likely indirect, for the same rate of interspecific sequence divergence is observed in all sequence subregions. A similar relationship between levels of recombination and neutral polymorphism in *Drosophila melanogaster* (Begun & Aquadro, 1992) has been explained in terms of the effects of genetic hitchhiking on linked genomic regions (interallelic exchange disrupts 'selective sweeps', resulting in higher levels of diversity in more recombinogenic regions). However, as discussed above, hitchhiking to advantageous mutants in malarious locales is unlikely to account for reduced diversity at the human ß-globin locus. Interspecific sequence differences also do not suggest the recent fixation of an amino acid substitution at the gene. An alternative hypothesis, which suggests that the continuous elimination of deleterious mutations from a selectively constrained genomic region can cause a concomitant reduction in the level of linked neutral polymorphism (Charlesworth, Morgan, & Charlesworth, 1993; Hudson, 1994), explains the lack of diversity in the ß-globin gene and 3' flanking region (and the

---

* Continued synthesis of HbF in adulthood can, by reducing the cellular concentration of HbS, ameliorate complications associated with sickle cell disease (Powars, 1991).

higher level of polymorphism in the 5' flanking DNA) better. Therefore, while positive Darwinian selection (balancing or directional) does not appear to have influenced the frequency or genomic distribution of silent variation at the ß-globin gene, the observed correlation between levels of variation and recombination is consistent with the action of purifying selection on the locus.

In the 5' flanking DNA, where proportionally more sequence variation was observed, an unusual clustering of polymorphic variation approximately 500 bp 5' of the ß-globin gene was also apparent. Mutations associated with the mechanics of interallelic exchange are unlikely to be responsible for the observed pattern: the initiation of recombination is presumed to occur considerably 5' of the sequenced region and the 3' resolution of observed conversion tracts is likely to be distance- not position-dependent (Chapter 3). There is, however, indirect evidence to suggest that the purine-pyrimidine (predominantly AT) repeat region where most of the point and length mutations occur is the initiation site for replication of a 200 kb region encompassing the human ß-globin gene cluster. The ß-globin origin of replication has been mapped to a 2 kb genomic region which includes the 5' flanking DNA sequenced here (Kitsberg et al., 1993). Within this 2 kb region is a conserved *PUR* sequence element which, when bound, is thought to promote duplex opening in adjacent AT-rich DNA (Bergemann & Johnson, 1992). Although not yet proven, it is likely that replication is initiated at such 'unwound' genomic segments. If so, the sequence differences observed here suggest that the initiation of replication is an error-prone process. In this context, it is interesting to note that two hypervariable regions in the mtDNA genome coincide with sequences associated either with the origin, or premature termination, of replication (Vigilant et al., 1991).

DNA sequence diversity at the human ß-globin gene demonstrates the extent to which selection (positive and purifying) and recombination can interact in their affect on both non-silent and silent polymorphic variation. However, as discussed in Chapter 3, the extensive and localised influence of recombination at the ß-globin gene is unlikely to be typical of that found at most nuclear loci, and it remains to be seen how the level and pattern of polymorphism observed here will compare to equivalent data from other

human genes. For instance, average sequence divergence at effectively silent sites in the gene and 3' flanking DNA, the least polymorphic genomic regions investigated here, is still twice that estimated on the basis of variation observed in human cDNA sequences (Li & Sadler, 1991). Therefore, the impact of recombination in these regions, while less pervasive than in the 5' flanking DNA, may nevertheless be sufficient to 'preserve' variation from losses associated with the effects of drift and purifying selection. A complete understanding of the role that recombination plays in determining underlying levels of neutral polymorphism, as well as the extent to which recombination influences the likelihood of observing genetic hitchhiking effects at human nuclear genes, will require DNA-level analysis of loci found in a variety of genomic and selective contexts.

Chapter 5

Geographic Differences in ß-Globin Diversity

Allelic sequence variation identified at the human ß-globin locus in the course of this research suggests that both the overall level of genetic diversity (as measured by the number of polymorphic sites relative to the total sample size, or $\hat{\theta}$) and the distribution of major types of sequence haplotypes (i.e. A, B, C, or D) varies from population to population. As we have also seen, however, neither molecular nor selective processes are likely to be wholly responsible for these apparent population differences: the impact of recombination (discussed in Chapter 3) is ubiquitous and there is no good evidence to suggest that malarial selection of ß haemoglobinopathies has influenced the level or genomic distribution of silent polymorphisms in affected locales (Chapter 4). ß-globin sequence diversity observed within and between geographic locales must therefore reflect predominantly the influence of demographic phenomena.

While there is little doubt that ß-globin diversity is a product of demographic influences, understanding precisely which processes have contributed to the variation observed in particular populations is far from straightforward. As all human populations derive ultimately from a single ancestral population, for example, genetic similarities observed among present-day populations could be assumed to reflect that shared common ancestry, and any differences attributed to the impact of forces (such as mutation, recombination, or random genetic drift) which have acted on descendant populations in the course of, or subsequent to, population dispersal. Although this relatively simple scenario is often valid, in other instances extensive gene flow can counteract genetic differentiation, so that certain populations remain more similar than they would be if they had evolved independently since divergence. Genetic drift can also, by chance, render the allelic composition of otherwise distantly-related populations indistinguishable. Because very different phenomena can result in the same general

outcome, multiple sources of information about the genetic relatedness of populations are required in order to decide which processes best explain genetic similarities and differences.

The five populations surveyed, Espiritu Santo, Maewo, the Eastern Highlands of Papua New Guinea or PNG, Palembang (in Sumatra), and the Gambia, permit the evaluation of genetic relationships in a variety of geographical and demographic contexts. Four of the five populations sampled, for example, are found in close proximity to one another in Southeast Asia (Indonesia and Melanesia) (Figure 5.1). Considerable linguistic and archaeological evidence (e.g. Bellwood, 1978; 1989; 1991) suggests a common origin of Southeast Asian populations, as well as extensive prehistoric population movements throughout Island Southeast Asia and Melanesia. It is likely, therefore, that the Indonesian and Melanesian populations surveyed here are genetically more similar to one another than to the African population. The genetic relationship between the Gambia and the other populations surveyed can be assessed in light of these expectations. Similarly, an *a priori* expectation of close genetic relatedness of the Melanesian populations (particularly Espiritu Santo and Maewo, neighbouring islands in the same archipelago), can be investigated in terms of the genetic composition of the relevant population samples. Unexpected genetic similarities and differences are likely to reflect the population-specific impact of processes such as gene flow or random genetic drift.

The nature of genetic similarities and differences among the five populations surveyed, measured in terms of the distribution of population-specific nucleotide and sequence haplotype polymorphisms as well as average pairwise sequence differences, were assessed according to such expected population relationships.

Testing the Extent of Population Genetic Differentiation

There are clear differences in the number and kind of sequence haplotypes identified in each of the populations surveyed (Chapter 3). The extent to which the observed

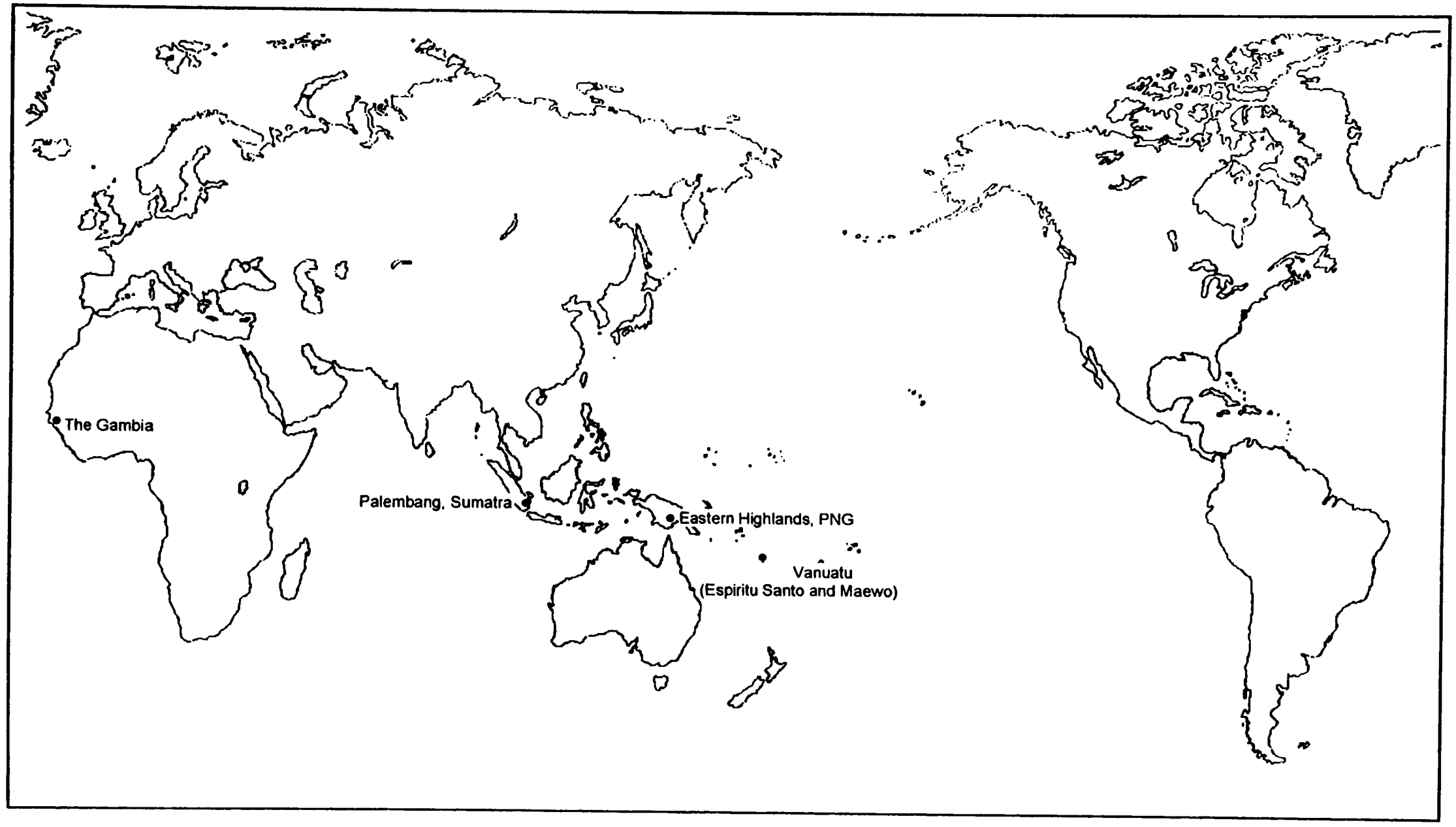


Figure 5.1 Map of the world, showing the geographic location of the five populations surveyed. PNG = Papua New Guinea.

sequence haplotype distributions result solely from the effects of sampling, or are instead due to substantive differences in the genetic composition of the locales investigated, may be determined by statistical assessment of observed and expected differences between the population samples. Allelic sequence diversity at the human ß-globin locus was evaluated in this way using the Hudson, Boos, and Kaplan (1992) test of geographic subdivision.

The Hudson, Boos, Kaplan Test of Geographic Subdivision

The Hudson, Boos, and Kaplan (HBK) test compares genetic differences observed between two samples with those that would be expected if the samples had been drawn from a single homogeneous population or gene pool. The expected amount of genetic differentiation is simulated by assigning each of the observed sequence haplotypes at random to two similarly-sized samples and determining the amount of divergence between them. The random assignment of sequence haplotypes is called a permutation and 1000 permutations are performed to generate a distribution of divergence estimates. If the observed estimate of genetic differentiation falls outside the 95% confidence interval of the randomly-generated distribution, the populations are considered significantly different genetically (Figure 5.2).

When DNA sequence data are available, the genetic differentiation of populations may be investigated in several different ways. One means of assessing genetic similarities and differences among populations is in terms of the frequencies of observed sequence haplotypes in each sample (an approach analogous to the analysis of allele frequency distributions using electrophoretic or restriction fragment length polymorphism data). However, if many sequence haplotypes are present in the populations surveyed, so that the number of sequences of each type is small, the power of the test is compromised. Another measure of genetic differentiation, which may be more appropriate for DNA sequence data, takes the number of pairwise nucleotide differences between sequence haplotypes, as well as their frequencies, into account. The HBK test evaluates geographic subdivision using both types of measures. It is, therefore, possible to compare the extent of differentiation among populations when different kinds of genetic

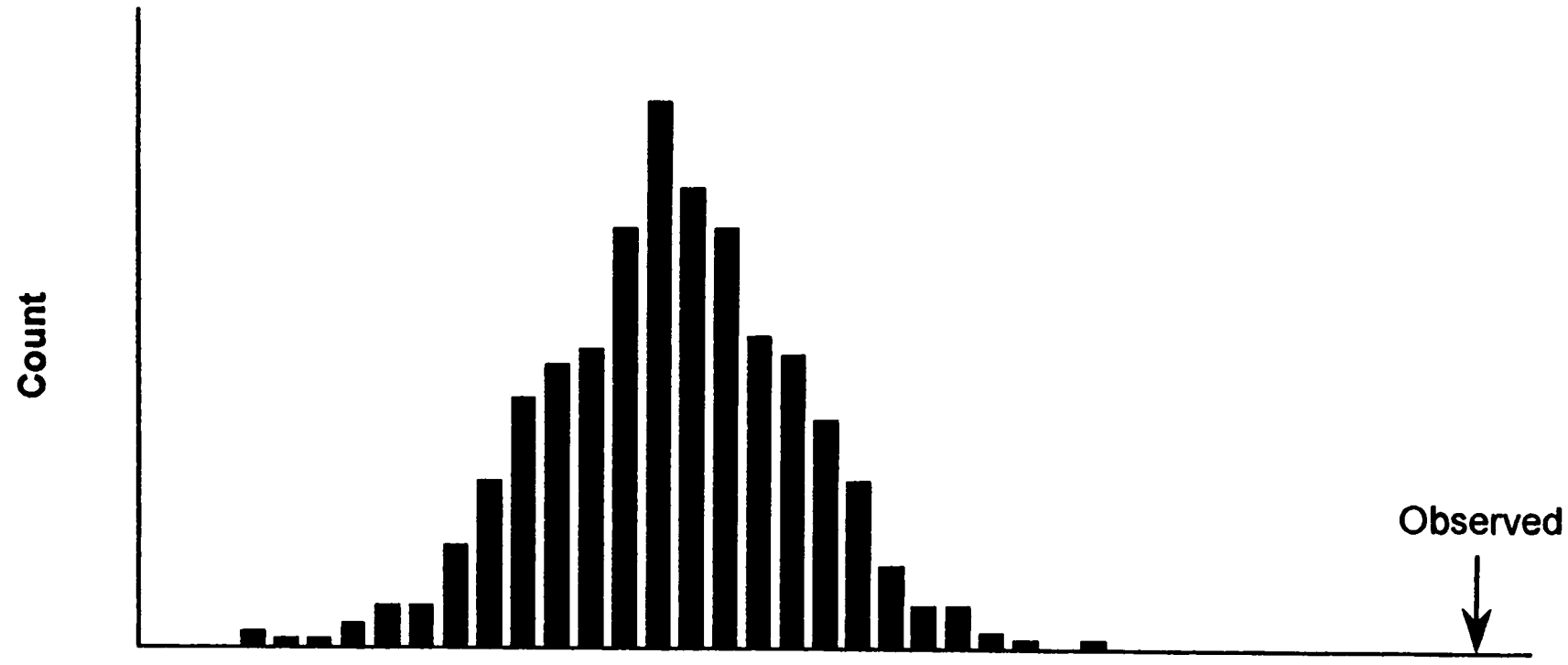


Figure 5.2 Hypothetical result of 1000 random permutations of sequences sampled from two subpopulations (based upon data presented by Stoneking et al., 1990). The observed value of Ks' for the comparison is significantly greater than any of the randomly-generated values. The probability, *P*, of observing such a value, given the null hypothesis that the populations are not genetically different, is less than 0.001.

variation are considered.

Genetic Differences Among the Population Samples

The probability values for each pairwise test of population genetic differentiation, using either a haplotype-based test statistic, χ^2 (above diagonal), or a pairwise difference statistic, K_S' (below diagonal), are given in Table 5.1 (these values were calculated using a computer program kindly provided by R. R. Hudson). When χ^2 was calculated, rare haplotypes were grouped at random ('lumped') so that the expected number of each haplotype in each locality was at least 1. This was done to assure that the numbers of each haplotype in each population were not too small. The K_S' measure used a logged estimate of nucleotide differences between sequence haplotypes to downweight sequence comparisons with large numbers of differences. Simulations performed by Hudson, Boos, and Kaplan (1992) suggest that these measures are the best (i.e. most powerful) statistics for the investigation geographic subdivision using their permutation-based test.

The estimated probability of genetic differentiation between samples taken from the islands of Espiritu Santo and Maewo in Vanuatu was 0.238 and 0.475 for the χ^2 and K_S' test statistics, respectively. As these results suggested that there were no genetic differences between the island populations, data from the two samples were combined into a single sample, Vanuatu, for the remainder of the test comparisons. The four resulting populations were significantly differentiated from one another, whether genetic differences were assessed using haplotype-based or pairwise difference statistics (Table 5.1). The only major discrepancy in the results obtained using the two types of measures was found for the PNG Highlands-Sumatra comparison, where the χ^2 test failed to detect a significant difference identified using the K_S' measure (0.010 for χ^2 versus 0.001 for K_S'). This difference may have arisen as a consequence of the lumping of rare haplotypes; many of the sequence haplotypes identified in PNG and Sumatra are unique.

Together, the results of the HBK test suggest that most of the locales investigated are

	Vanuatu (n=61)	PNG Highlands (n=24)	Sumatra (n=28)	Gambia (n=31)
Vanuatu		0.000	0.001	0.000
Highlands	0.000		0.010	0.000
Sumatra	0.002	0.001		0.000
Gambia	0.000	0.000	0.000	

Table 5.1 Estimated *P* values, based on 1000 random permutations (Hudson, Boos, & Kaplan, 1992), for the $\chi 2$ (above diagonal) and Ks' (below diagonal) test statistics.

genetically distinct, at least when variation at the ß-globin locus is considered. Indeed, the only samples which appear to be drawn from the same gene pool are those taken from nearby islands in Vanuatu. Although the evidence for geographic subdivision is unequivocal, it is very difficult to tell what processes have contributed to the inferred differentiation. To gain insight into the evolutionary processes which might be influencing the genetic differentiation of the geographic locales surveyed, the distribution of nucleotide and sequence haplotype diversity among the populations was investigated.

Distribution of Sequence Polymorphism Among Geographic Locales

The way in which ß-globin allelic sequence diversity is apportioned within and between surveyed populations can be investigated in both qualitative and quantitative terms. The distribution of population-specific nucleotide and sequence haplotype polymorphism, which is described below, provides a description of the extent to which variation is confined to particular populations. This description is then augmented by a numerical summary of the data, in which genetic similarities and differences are defined in terms of the net sequence divergence between locales.

Population-Specific Nucleotide and Sequence Haplotype Diversity

The number of population-specific nucleotide polymorphisms and sequence haplotypes observed in each of the four populations (data from Espiritu Santo and Maewo considered together as Vanuatu) is presented in Table 5.2. As shown, only a portion of the nucleotide polymorphism observed in each locale was population-specific. The largest proportion of population-specific site polymorphism was observed in the Gambia: 11 (41%) of the 27 variable sites were population-specific. In contrast, only approximately 5% of the variable sites identified in Vanuatu, the PNG Highlands, and Sumatra were unique (1 out of 20, 16, 18 sites respectively). The small number of population-specific differences observed in the non-African samples, which derive from locales in Southeast Asia and Melanesia, may reflect the comparative similarity of

	By Population				By Major Geographic Locale		
	Vanuatu (n=61)	PNG Highlands (n=24)	Sumatra (n=28)	Gambia (n=31)	SE Asia (Melanesia)* (n=113)	Africa (n=31)	All / Other†
Nucleotide Polymorphisms							
Number Unique	1	1	1	11	7 (3)	11	13 / 3
Total Polymorphic	20	16	18	27	23 (22)	27	34
Percent of Total Unique	5	6.3	5.5	41	30.4 (13.6)	40.7	38.2 / 8.8
Sequence Haplotypes							
Number Unique	5	4	9	21	30 (12)	21	1 / 1
Total Polymorphic	18	9	20	23	32 (22)	23	53
Percent of Total Unique	27.7	44.4	45	91.3	93.8 (54.5)	91.3	1.9 / 1.9

* That portion of SE Asian variation which is confined to Vanuatu and the PNG Highlands is given in parentheses.

† The number of nucleotide polymorphisms and/or sequence haplotypes found in all populations (All) versus those observed in certain Asian and African populations (Other).

Table 5.2 The proportion of population-specific nucleotide polymorphisms, and sequence haplotypes, observed in each population and major geographic region.

populations which lie in closer proximity to one another than to the Gambia. When examined collectively, 7 (30%) of the 23 polymorphic sites observed in the three non-African populations were Asian-specific, a figure roughly comparable to the proportion of African-specific polymorphisms observed in the Gambia. Approximately one-half (3 out of 7) of the Asian-specific nucleotide polymorphisms were found exclusively in Melanesia (i.e. Vanuatu and the PNG Highlands).

As described in Chapter 3, 34 nucleotide polymorphisms were identified in the total sample. Seven of these sites (21%) were observed only in Southeast Asian (Indonesian and Melanesian) populations, 11 (32%) were confined to the African (Gambian) population, and 13 (38%) were found in all four populations surveyed. Three remaining polymorphisms (9%) were shared between certain Southeast Asian populations and the Gambia. Therefore, although a large proportion of variable sites are shared among populations, considerable differences exist between African and Asian locales.

A similar comparison of sequence haplotype polymorphism among the four populations surveyed is also shown in Table 5.2. Again, only a portion of observed haplotype polymorphism was confined to particular populations and, as observed for polymorphic sites, the largest proportion of unique sequence haplotypes was found in the sample from the Gambia. However, more population-specific polymorphism was observed in each of the locales surveyed when sequence haplotype diversity was compared: 28, 44, and 45% of the sequence haplotypes observed in Vanuatu, the PNG Highlands, and Sumatra, respectively, were unique to each population, and 91% (21) of the 23 sequence haplotypes observed in the Gambia were population-specific.

When polymorphism in the three Asian populations was considered collectively, 94% (30) of the 32 sequence haplotypes observed were Asian-specific. Thus, as seen with the nucleotide polymorphisms, a similar proportion of sequence haplotypes were unique to Asian and African populations. The overall distribution of sequence haplotypes within and between major geographic locales was different, however: 57% of the 53 sequence haplotypes identified in the total sample were confined to Southeast Asia and 40% to Africa, while only 1 sequence haplotype (2%) was shared

among all five populations. The greater differentiation of major geographic locales (and populations) at the sequence haplotype level suggests that new ß-globin alleles arise more often from the effects of interallelic recombination than point mutation. This finding is consistent with the high rate of recombination inferred to occur in sequences 5' of the ß-globin gene (Chapter 3).

Net Number of Sequence Differences Between Populations

The observed distribution of population-specific nucleotide and sequence haplotype diversity at the human ß-globin locus suggests that major genetic distinctions exist between African and Asian populations and that geographic differentiation is most extensive when sequence haplotype variation is considered. In order to investigate further genetic similarities and differences between pairs of populations, a numerical estimate of genetic divergence is required. A suitable measure, which incorporates information about nucleotide *and* sequence haplotype polymorphism (in much the same manner as the K_S' test statistic discussed above), is the net number of sequence differences between populations.

The net number of sequence differences between populations *i* and *j* can be estimated from the average pairwise sequence differences observed within and between the populations (Nei & Li, 1979):

$$\hat{d}_{ij} = \hat{\pi}_{ij} - \frac{\hat{\pi}_i + \hat{\pi}_j}{2} \qquad (5.1)$$

$\hat{\pi}_{ij}$ is the average pairwise difference *between* sequences sampled from populations *i* and *j*, estimated by

$$\hat{\pi}_{ij} = \sum_{xy} p_{ix} p_{jy} \pi_{xy} \qquad (5.2)$$

where p_{ix} is equal to the estimated frequency of haplotype *x* in population *i* and π_{xy} is the number of nucleotide (and/or length) differences observed between the *x* and *y*

sequence haplotypes. $\hat{\pi}_i$ is the average pairwise difference between sequences sampled from *within* population *i* only, estimated by

$$\hat{\pi}_i = \frac{n_i}{n_i - 1} \sum_{xy} p_{ix} p_{iy} \pi_{xy} \qquad (5.3)$$

where n_i is the total number of sequences surveyed in population *i*. The net number of sequence differences therefore quantifies the amount of genetic difference which remains between locales after intrapopulation sequence diversity is taken into account.

Within- (on the diagonal, in bold) and between- (above the diagonal) population average pairwise differences, and the resulting number of net sequence differences between populations (below the diagonal), are presented in Table 5.3. The average number of differences between sequence haplotypes from the same population ranged from 5.97 in the PNG Highlands to 6.58 in Sumatra (mean, 6.21). In other words, approximately 1 in 430 basepairs (bp) vary between any two randomly chosen alleles sampled from the same population. The average number of differences between sequences drawn from different populations was higher, in every instance, than the within-population estimates (range 6.86 to 7.96, mean 7.37), consistent with the genetic differentiation of populations inferred from the HBK test. The net number of sequence differences between each pair of populations was therefore positive and ranged from 0.72 (between Vanuatu and the Gambia) to 1.84 (between the PNG Highlands and the Gambia), averaging 1.16 over all comparisons. Thus, less than 20% of the variation observed within each population contributed to between-population genetic differences.

Individual estimates of net sequence divergence suggest that there is considerable variation in the extent to which ß-globin diversity differs between populations, consistent with the distribution of population-specific polymorphisms discussed above. However, because the estimate of the net number of sequence differences is determined by similarities, as well as differences, in the sequence haplotype distributions of particular populations, the resulting *pattern* of population divergence is not the same as that

	Vanuatu (n=61)	PNG Highlands (n=24)	Sumatra (n=28)	Gambia (n=31)
Vanuatu	**6.03**	7.00	7.24	6.86
Highlands	1.00	**5.97**	7.22	7.96
Sumatra	0.94	0.95	**6.58**	7.93
Gambia	0.72	1.84	1.51	**6.26**

Table 5.3 Summary of average pairwise sequence differences within- (diagonal) and between- (above diagonal) populations. Net sequence differences between populations are given below the diagonal.

suggested by population-specific sequence differences. A significant proportion of population-specific nucleotide and sequence haplotype polymorphisms distinguished the Gambia from the other three populations surveyed, for instance, which were themselves only moderately differentiated. When net sequence differences were considered, however, the Gambia was strongly differentiated from PNG (1.84) and Sumatra (1.51), but not from Vanuatu (0.71), and more differences were found between Vanuatu and the PNG Highlands (1.00), than between either Melanesian population and Sumatra (0.94 and 0.95, respectively).

While the HBK test suggested no more than the fact that equally significant differences in ß-globin diversity exist between each of the populations surveyed, the distribution of allelic sequence diversity within and between populations demonstrates that some geographic locales are more genetically-distinct than others. Qualitative and quantitative descriptions of the data differ in terms of the genetic distances that each imply. The net number of sequence differences between populations appear to be more a function of genetic similarities between samples than a reflection of the number of population-specific differences. It is not immediately apparent, however, why net sequence differences between Island Melanesia and the Gambia are smaller than between Vanuatu and the PNG Highlands - two ethnically-related, and geographically proximate, Melanesian populations. To explore further the molecular and demographic processes which have contributed to these patterns of differentiation, overall genetic relationships among the sampled populations, taking into consideration patterns of variation in different genomic subregions, were investigated.

The Basis of Geographic Differences in Diversity at the ß-Globin Locus

As there is no evidence for the effects of natural selection on polymorphism at the ß-globin locus, geographic differences in allelic sequence diversity are likely to derive from the combined influence of molecular and demographic phenomena. Given the high rate of interallelic recombination in the genomic region 5' of the ß-globin gene, in particular, and its inferred (indirect) effect on levels of polymorphism in that region

(Chapter 4), it is important to investigate whether recombination may interact with demographic processes in such a way as to determine the geographic distribution of polymorphism at the locus. Therefore, net sequence differences among the populations were recalculated separately for the 5' flanking and gene + 3' flanking regions respectively.

The Genomic Distribution of Geographic Differentiation

Estimates of within- and between-population average pairwise differences, and resulting estimates of net sequence divergence, were calculated with respect to (i.e. divided by) the total number of effectively silent sites surveyed in the 5' flanking and 3' gene + flanking subregions of the 3 kb ß-globin sequence (863 and 1795 bp respectively). This was done so that the distribution of sequence differences in the differently-sized regions could be compared directly. Per nucleotide estimates for the 5' and 3' regions, as well as the total sequence, are presented in Table 5.4. In the 5' flanking DNA (where proportionately more recombination occurs) within- and between-population average pairwise sequence differences were, on average, over two times larger than those observed in the gene and 3' flanking region (in accord with the higher overall level of polymorphism observed in that region, as discussed in Chapter 4). These differences were reflected in the net number of sequence differences between populations estimated for each subregion.

Besides being considerably larger, however, estimates of net sequence divergence in the 5' flanking DNA also suggested a different pattern of population divergence from that found for the 3' portion of the ß-globin sequence. For instance, in the 5' flanking DNA, the largest genetic distances were observed between the PNG Highlands and the Gambia, and between the PNG Highlands and Sumatra (0.144% and 0.096% differences per nucleotide, respectively), whereas in the 3' subregion, those pairs of populations were the second and fifth (out of six) most divergent, and Sumatra and the Gambia had the largest number of net sequence differences instead (0.52% per nucleotide). Similarly, the smallest number of net sequence differences in the 5' subregion were observed between Vanuatu and Sumatra (0.044%), which were the

	Vanuatu	Highlands	Sumatra	Gambia
Total Sequence				
Vanuatu	**0.227%**	0.263%	0.272%	0.258%
Highlands	0.038%	**0.225%**	0.272%	0.299%
Sumatra	0.035%	0.036%	**0.247%**	0.298%
Gambia	0.027%	0.069%	0.057%	**0.236%**
5' Flanking DNA				
Vanuatu	**0.319%**	0.411%	0.399%	0.454%
Highlands	0.069%	**0.364%**	0.474%	0.548%
Sumatra	0.044%	0.096%	**0.392%**	0.484%
Gambia	0.073%	0.144%	0.066%	**0.444%**
Gene + 3' Flanking DNA				
Vanuatu	**0.182%**	0.192%	0.212%	0.148%
Highlands	0.023%	**0.157%**	0.174%	0.180%
Sumatra	0.031%	0.007%	**0.178%**	0.209%
Gambia	0.005%	0.034%	0.052%	**0.136%**

Table 5.4 Summary of average pairwise sequence difference within- (diagonal) and between- (above diagonal) populations, and net sequence differences between populations (below diagonal), for 5' and 3' subregions of the ß-globin sequence. Values represent per nucleotide estimates of sequence divergence. The greatest net sequence difference for each subregion is boxed, the smallest underlined.

third most divergent pair of populations in the 3' half of the ß-globin locus; considerably fewer net sequence differences were observed between Vanuatu and the Gambia, and Sumatra and the PNG Highlands, in the gene and 3' flanking DNA.

As discussed in Chapter 3, ß-globin allelic sequence diversity is characterised by the presence of four groups of sequence haplotypes, defined by the pattern of variation found in the gene and the 3' flanking DNA, which are found in varying proportions in each of the surveyed populations (Table 3.3). In Vanuatu and the Gambia, A- and B-type sequences predominate, whereas in the PNG Highlands and Sumatra, the majority of the sequences identified belong to the B and C sequence haplotype groups. The small number of net sequence differences observed between these pairs of populations at the 3' end of the ß-globin region therefore appear to derive from their similar overall sequence haplotype distributions. To confirm this supposition, genetic differences between populations were re-evaluated, using the Chord distance measure (Edwards, 1971) solely in terms of the observed frequencies of the four types of sequence haplotype in each sample. These genetic distances were significantly correlated with 3' net sequence divergence estimates (Mantel matrix correlation coefficient, r = 0.85, *P* = 0.04 - personal communication, R. M. Harding), suggesting that the major types of sequence haplotype found in each sample determine the pattern of population divergence identified in the 3' genomic region.

While there is linguistic evidence (Würm, 1983) to suggest that the movement of languages (and possibly people) from Indonesia to Papua New Guinea has occurred in the last five to ten thousand years (migration which, if accompanied by gene flow, might explain similarities in the distribution of ß-globin sequence haplotypes in the two locations), extensive migration between Melanesia and Africa is unlikely to explain the large number of A- and B-type sequence haplotypes observed in both Vanuatu and the Gambia. The latter two populations are ethnically and linguistically distinct, and additional, 3' flanking region, variation observed among group B sequences in the Gambia (Chapter 3) is not found in Island Melanesia. Instead, similarities in the populations' haplotype distributions are most likely coincidental, arising either by chance as a consequence of the random sampling of gametes during zygote formation

(i.e. random genetic drift), or as a consequence of gene flow into one or both populations from other geographic regions. The influence of drift and/or geographic isolation on patterns of population differentiation was explored using an overall measure of genetic differentiation among populations.

Differentiation of Populations Relative to Variation in the Total Sample

The extent to which genetic variation in populations can change as a consequence of drift, mutation, or recombination, depends on the number of breeding individuals in the population, or effective population size, and the extent of geographic isolation from other populations. Both population size and migration determine the extent to which sequence diversity will be affected by random genetic drift. Drift is greatest in small-sized populations, and migration counteracts the effects of drift by constantly introducing new variants to the population. Gene flow among geographic regions can also determine the degree to which population-specific nucleotide and sequence haplotype polymorphisms accumulate. When the rate of mutation/recombination is high in relation to migration, differences between populations are large, and vice-versa. Population genetic differentiation, particularly in relation to variation found in the population as a whole (or in this case, the total sample investigated), can therefore provide important insights into such phenomena by providing an indication of the relative genetic isolation of populations.

The probability of identity of randomly chosen sequence haplotypes drawn from a single population relative to sequence haplotypes drawn from the population as a whole can be described using the population fixation index, F_{ST} (Wright, 1951). When DNA sequence data are available, F_{ST} is estimated from average pairwise sequence differences within and between populations, and designated N_{ST} (because the estimate is based on genetic divergence at the nucleotide level). N_{ST} is calculated as the ratio of the average genetic distance between genes from different populations, $\hat{v}_b$, relative to that among genes in the population at large (i.e. the sum of within-, $\hat{v}_w$, and between-population differences) (Lynch & Crease, 1990):

$$N_{ST} = \frac{\hat{v}_b}{\hat{v}_w + \hat{v}_b} \tag{5.4}$$

where

$$\hat{v}_w = \frac{\sum \hat{\pi}_i}{n_p} \tag{5.4a}$$

$$\hat{v}_b = \frac{2\sum \hat{d}_{ij}}{n_p(n_p - 1)} \tag{5.4b}$$

n_p is the total number of populations under consideration, and $\hat{\pi}_i$ and $\hat{d}_{ij}$ are the average number of pairwise differences between sequences from within population *i* and the net number of sequence differences between populations *i* and *j*, respectively, as defined by Equations 5.3 and 5.1. N_{ST} estimates range from 0, if there are no differences between populations, to 1, for the case of complete genetic differentiation.

Per nucleotide values of $\hat{v}_w$ and $\hat{v}_b$, and the resulting estimate of N_{ST}, for the total sample (n=144), as well as four subsets of the total sample, generated by removing each of the populations in turn (after Slatkin, 1985 and Stoneking et al., 1990), are given in Table 5.5; all estimates were derived for variation in the total 3 kb sequence, as well as the 5' and 3' subregions of the ß-globin locus. As shown, when the entire 3 kb region is considered, N_{ST} for the total sample is 0.157. This suggests that 16% of the total variation observed at the ß-globin locus occurs between particular geographic locales while the remaining 84% is found within populations (a result analogous to the more general finding that net sequence differences among pairs of populations were, on average, less than one-fifth the number of average pairwise differences observed within populations). Similar estimates of N_{ST} for the 5' and 3' subregions are 0.177 and 0.134, respectively. Thus, as described above, more sequence differences are found between populations in the 5' flanking DNA, than in the gene and 3' flanking regions.

The extent to which each of the populations is isolated genetically with respect to other populations is indicated by estimates of N_{ST} calculated for the four subsets of the total

Population Removed	5'	3'	TOTAL
None (All 4 Populations)			
v_w	0.380%	0.163%	0.234%
v_b	0.082%	0.025%	0.044%
N_{ST}	0.177	0.134	0.157
Vanuatu			
v_w	0.400%	0.157%	0.236%
v_b	0.102%	0.031%	0.054%
N_{ST}	0.203	0.164	0.186
PNG Highlands			
v_w	0.385%	0.165%	0.237%
v_b	0.061%	0.030%	0.040%
N_{ST}	<u>0.137</u>	0.152	<u>0.144</u>
Sumatra			
v_w	0.376%	0.158%	0.229%
v_b	0.095%	0.020%	0.045%
N_{ST}	0.202	<u>0.114</u>	0.163
Gambia			
v_w	0.358%	0.173%	0.233%
v_b	0.070%	0.020%	0.036%
N_{ST}	<u>0.162</u>	<u>0.105</u>	<u>0.135</u>

Table 5.5 Estimated N_{ST} values for the human ß-globin locus, removing each population in turn. v_w and v_b are the within- and between-population components of N_{ST}, respectively, given here as per nucleotide estimates. Values of N_{ST} lower than that estimated for the total sample are underlined.

sample. For the entire 3 kb sequence, N_{ST} estimates were the same or higher than that calculated for the total sample when either Vanuatu or Sumatra were removed (Table 5.5). When variation in the PNG Highlands and the Gambia was disregarded, however, N_{ST} was reduced by 8.3% (0.144) and 14% (0.135), respectively, suggesting that variation in both locales is, to a certain extent, distinct from that observed in other populations. Estimates of N_{ST} for the 5' flanking DNA were also consistent with the genetic isolation of both the PNG Highlands and the Gambia, although in this instance, removing the PNG Highlands made a greater difference to the estimate of N_{ST} (reduced by 22.6%) than removing the Gambia (8.5% lower). In contrast, when variation in the gene and 3' flanking DNA was considered, the comparative isolation of the PNG Highlands population was not indicated. Instead, removing Sumatra and the Gambia from the total sample reduced estimates of N_{ST} by 14.9 and 21.6%, respectively.

Taken collectively, these results suggest that variation in the Gambia, in particular, is consistent with the geographic isolation of the African population from the other populations surveyed. As described earlier, a considerable proportion of the nucleotide and sequence haplotype polymorphism observed in the Gambia is population-specific, and it is these differences, which are removed when the Gambian sample is disregarded, which result in a lower overall estimate of N_{ST} for the remaining populations. In the 3' half of the ß-globin sequence, where recombination is less common and the effects of purifying selection and/or random genetic drift presumably more important (Chapter 4), relative isolation of the Sumatran population is also indicated. Given the geographic proximity (and presumed recent evolutionary origin) of populations living in PNG and Island Melanesia, this result is not totally unexpected. The effect of the exclusion of PNG Highland variation on the geographical distribution of polymorphism found in the 5' flanking DNA (as well as in the entire 3 kb sequence), which is not consistent with the results obtained for the 3' subregion, suggests that something other than geographic isolation has contributed to the genetic differentiation of that population.

The Structure of ß-Globin Sequence Variation

As these analyses clearly demonstrate, significant genetic differences exist among four of the five populations surveyed. Population-specific nucleotide polymorphisms and sequence haplotypes were observed in every locale except Maewo (the sample from Maewo was genetically-indistinguishable from the Espiritu Santo sample). Average pairwise differences between sequence haplotypes drawn from different populations were higher in all cases than those found in the same population. The populations were not, however, all equally divergent. Rather, some locales shared more genetic similarities than others, as reflected in the net number of sequence differences between alleles drawn at random from particular pairs of samples. This measure unexpectedly suggested that certain geographically-distant populations (such as Vanuatu and the Gambia) were genetically more similar to one another than to neighbouring locales, although when relationships among all five populations were considered, the African sample was clearly differentiated with respect to the Southeast Asian populations. Furthermore, the pattern of population relationship suggested by sequence differences in the 5' flanking region differed from that found when only the gene and 3' flanking regions were considered. The lack of evidence for the effects of selection on the locus (Chapter 4) mean that the observed structure of ß-globin variation (the distribution of allelic sequence differences within and between populations) must derive from the effects of other evolutionary phenomena.

The distribution of neutral variation within and between populations is subject to the effects of random genetic drift, migration (when it involves the movement of alleles, or gene flow), mutation, and interallelic recombination. In the absence of mutation and recombination, completely isolated populations can accumulate genetic changes simply as a consequence of drift (Wright 1931), through the loss of low-frequency alleles or as a result of changes to the overall allele frequency distribution. More pronounced changes are expected in small populations because the effects of drift are proportionately greater. Mutation and recombination accentuate this underlying process of differentiation by the regular introduction of new allelic variants to specific locales. When gene flow is limited or non-existent, new alleles remain confined to the

populations in which they arose, and combine with differences resulting from drift, to produce population-specific genetic profiles. Therefore, at a neutrally evolving locus, geographic differences in the level and pattern of polymorphism reflect the amount of time that particular populations have been evolving independently, as well as the level of gene flow which has occurred between them since they diverged.

As gene flow is expected to counteract the differentiating effects of mutation, recombination, and genetic drift, it is important to consider whether the structure of the observed ß-globin variation is consistent with extensive migration between the sampled locales. Although human migratory distances are, on average, quite small (Wijsman & Cavalli-Sforza, 1984), humans can and do migrate hundreds or thousands of mile from their place of birth. It is therefore unlikely that the four populations investigated here have shared no migrants in the recent past. Nevertheless, the limited distribution of a significant proportion of the nucleotide polymorphisms identified suggests that gene flow is not so extensive that newly generated polymorphisms are immediately disseminated to surrounding locales. The greatest number of population-specific nucleotide polymorphisms distinguish Asian and African populations, in accord with the considerable geographic distance which exists between the Gambia and the other locales surveyed. Even among the more closely situated Southeast Asian populations, however, many population-specific sequence haplotypes were identified, suggesting that the rate of creation of new alleles at the locus (by recombination in particular) is higher than the level of interpopulation gene flow.

Despite the apparently low level of gene flow suggested by the observed population-specific polymorphism, the four locales surveyed were not highly differentiated: only 16% of the total variation identified at the ß-globin locus represented sequence differences found *between* populations. Of course even small levels of gene flow will, over long periods, introduce a given polymorphism to most geographic regions. If populations are at equilibrium with respect to the effects of mutation/recombination and drift, and the rate of introduction of new alleles is only slightly greater than the level of gene flow, limited differentiation (such as that observed here) may result. The nature of the sequence haplotypes observed in each population, however, suggests that

genetic similarities among the sampled locales are more likely to derive from recent common ancestry than long-term gene flow. In particular, the presence of four distinct types of ß-globin gene (but no 'intermediate' sequence haplotypes) in each of the populations surveyed is difficult to reconcile with the gradual, and presumably stochastic, dispersal of alleles which have originated in distinct geographic locations.

The genetic similarities observed among the four locales, if largely a product of common ancestry, suggest that the populations have only been evolving independently for a short period of time (discussed in greater detail in Chapter 6). Given the apparent recency of population divergence, it is likely that the observed genetic differentiation derives more from the on-going impact of mutation and recombination than with changes in the nature and/or frequency of alleles resulting from the effects of random genetic drift. A possible exception, however, is highlighted by the unusual genetic relationship of Vanuatu to the other locales investigated: the net number of sequence differences between alleles observed in Vanuatu and the Gambia was considerably smaller than those found between Vanuatu and either the PNG Highlands or Sumatra, despite the extensive differentiation of African and Asian locales suggested by population-specific polymorphisms. The simplest explanation for such an unexpected finding is drift in Vanuatu, which has changed the allelic composition of the population from the Southeast Asian pattern (i.e. that which is observed in the PNG Highlands and Sumatra) to one which coincidentally resembles the Gambian sequence haplotype distribution.

However, there is no other evidence to suggest the impact of drift on ß-globin variation in Vanuatu. Neither the overall level of nucleotide diversity in the population (Chapter 4), nor the observed number of sequence haplotypes, was lower than that found in the other Southeast Asian populations investigated; a reduction in diversity might be expected if a population bottleneck associated with the colonisation of the island archipelago (or any subsequent reduction in effective population size) was responsible for the observed shift in allelic composition (Tajima, 1989b). Rather, whatever evolutionary process has affected the proportions of particular types of alleles in Vanuatu has done so without having had any effect on the overall level of genetic

diversity in the population. Although random genetic drift in a constant-sized and genetically-isolated population could, over a long time period, result in such changes, the available evidence does not suggest that Vanuatu is isolated with respect to the other populations surveyed: estimates of N_{ST} for the total sample (5' , 3', or total sequence), for instance, were not reduced when Vanuatu was excluded from consideration.

Of the three major Southeast Asian populations investigated in the present study, only the PNG Highlands sample was found to be isolated with respect to other locales. The unexpected differentiation of that population, in contrast to the genetic isolation of the geographically-distant Gambian population, is more likely to have resulted from a reduction in size of the PNG Highlands population during or after population divergence than from restricted gene flow. Less clear is the extent to which the observed genetic differences arise as a consequence of drift. The low overall level of diversity in the PNG Highlands sample, as measured by the average number of pairwise differences among observed sequence haplotypes, is consistent with the increased impact of drift in a smaller-sized population. However, polymorphism in only a limited portion of the 3 kb region (the 800 bp 5' of the ß-globin gene) actually suggests PNG Highland genetic isolation. The presence of such genomic heterogeneity makes it unlikely that drift is the primary determinant of PNG Highland differentiation. Instead, a concomitant reduction in interallelic recombination, sufficient to render variation in the normally recombinogenic 5' flanking region susceptible to the effects of purifying selection (Chapter 4), may be a more plausible explanation for the observed patterns.

Together, the available data suggest that neither recent common ancestry nor random genetic drift is likely to be responsible for the concordant sequence haplotype distributions observed in Vanuatu and the Gambia. As it is also unlikely that mutation and recombination have contributed to the unexpected allelic composition of the Vanuatuan sample, the influence of gene flow is strongly implicated. In fact, archaeological and linguistic evidence suggest the introduction to Island Melanesia (but not Highland PNG or Sumatra) of Austronesian-speaking populations beginning 3500 years ago (Bellwood, 1989). Material remains associated with these migrants, in

particular a distinctive form of decorative pottery called 'Lapita', trace the rapid expansion of the Austronesians through the Solomon Islands and Vanuatu and eastward into the central Pacific (Diamond, 1988). The presence of Melanesian genotypes in present-day Polynesian populations (Hertzberg, Mickleson, & Trent, 1988; Hill, O'Shaughnessy, & Clegg, 1989) confirms that substantial amounts of gene flow occurred between Island Melanesian and Austronesian populations during the course of that expansion. It is therefore possible that gene flow associated with the Austronesian dispersal may have contributed to the unusual ß-globin sequence haplotype distribution observed in Vanuatu (no firm conclusions can be drawn, however, until further investigation of Southeast Asian and Polynesian populations is undertaken).

Geographic differences in ß-globin diversity reflect, to a large extent, the overall genetic structure of human populations. The estimate of N_{ST} suggested by polymorphisms found in the non-recombining portions of the 3 kb ß-globin gene region (N_{ST} = 0.134), for instance, is almost identical to an estimate calculated on the basis of the frequencies of 100 nuclear DNA restriction fragment length polymorphisms (RFLPs) found in African, European, Asian, and Melanesian populations (F_{ST} = 0.137 ± 0.010, Bowcock et al., 1991a; 1991b). Whereas it has proven exceedingly difficult to interpret RFLP variation in terms of the effects of processes such as drift and admixture, fine-scale investigation of allelic sequence polymorphism at the human ß-globin locus has provided important insights into the nature of interpopulation gene flow, the extent to which mutation and recombination contribute to population differentiation, and the role of drift in determining genetic change. ß-globin variation suggests that recent common ancestry best explains the extensive genetic similarities which exist among most human populations and that mutation and recombination, in closely-related but otherwise subdivided populations, may be more important to geographic differentiation than random genetic drift. The relationship of observed geographic differences to human evolutionary history is discussed in Chapter 6.

Chapter 6

ß-Globin Sequence Diversity and Human Evolution

This investigation of DNA sequence diversity at the human nuclear locus, ß-globin, has identified considerable nucleotide and sequence haplotype polymorphism - genetic variation which reflects the impact of both molecular and demographic processes. Thirty-four polymorphic sites (two resulting in ß haemoglobinopathies, the remainder phenotypically-silent) were found to vary in the 3 kilobase (kb) region examined. The genomic distribution of these sites was non-uniform, with a disproportionately large number of polymorphisms (including several small insertions and deletions) being found in a region approximately 500 basepairs (bp) 5' of the ß-globin gene. All but one silent polymorphic site was observed to occur in non-coding (intervening and/or flanking) DNA. The bulk of this highly-informative genetic variability would therefore have remained undetected if protein electrophoresis or restriction fragment length polymorphism (RFLP), rather than DNA sequence, analysis had been employed.

In the five populations surveyed, approximately 1 in 500 bp differed on average between any two randomly-chosen ß-globin alleles. The nature of the nucleotide changes present on particular chromosomes, described, collectively, in terms of allelic sequence haplotypes, was determined by direct sequence analysis of allele-specific PCR amplification products. Fifty-three unique sequence haplotypes were identified among the 173 chromosomes investigated. These sequence haplotypes fell into four major groups, designated A, B, C, or D, based on conserved sequence similarities at the gene and 3' flanking DNA (sequence haplotype subgroups were also apparent). Most of the variation within each of these major haplotype classes was shown to derive from the impact of interallelic recombination, whose effects were restricted primarily to 5' flanking region DNA. The high level of inferred recombinatorial activity probably relates to the presence of a previously-identified recombination hotspot 5' of the ß-

globin gene.

Both the observed level of nucleotide diversity, and the inferred amount of 'excess' haplotype polymorphism resulting from recombination, varied among the populations surveyed, suggesting that population-specific processes have also had an important influence on DNA sequence variation at the human ß-globin gene. Such processes have contributed to the significant genetic differentiation of the populations: nearly one-half of the observed nucleotide polymorphisms, and three-quarters of all identified sequence haplotypes, were found to be confined to specific populations. In addition, the observed frequencies of major types of sequence haplotype differed considerably among sampled locales. It is unlikely that selection is responsible for these differences because, as shown, observed patterns of nucleotide and sequence haplotype polymorphism are not inconsistent with neutral theory expectations. Rather, geographic differences in ß-globin diversity were found to reflect differences in the effective population size of, and patterns of migration among, the semi-isolated populations.

The apparently neutral nature of ß-globin allelic sequence diversity means that it is possible to interpret the genetic variation observed within and between populations, as well as major geographic regions, in terms of the nature and timing of population differentiation (for example, populations which share the fewest sequence similarities are likely to have been diverging for the longest periods of time). Inferences derived from the populations investigated here are, of course, relevant to larger human population relationships as well. The implications of observed ß-globin sequence variation for our understanding of human evolutionary origins, which are at present only incompletely understood, are discussed here.

Human Evolution: Current Hypotheses

Fossil remains suggest that, 2 million or more years ago (Swisher et al., 1994), a pre-human hominid species, *Homo erectus* (or similar early *Homo* species, see Gibbons, 1994) dispersed from its continent of origin, Africa, to much of the Old World. In the

next 1 to 2 million years, anatomically modern human beings (*Homo sapiens sapiens*) evolved from this precursor species, so that by approximately 50,000 years ago, all remaining *erectus* populations had been replaced by fully modern humans. Precisely when and where *H. s. sapiens* originated from *H. erectus* is less certain, and hotly contested (see Stringer & Bräuer, 1994 and Frayer et al., 1994 for the latest chapter in this on-going, and increasingly vociferous, debate). One theory, known as the 'multiregional' hypothesis (Wolpoff, 1989; Frayer et al., 1994) suggests that *H. s. sapiens* evolved *in situ* from geographically-dispersed *H. erectus* populations. Other palaeontologists (Stringer & Andrews, 1988; Stringer, 1992; Stringer & Bräuer, 1994) believe the evidence suggests that the transition to anatomically modern humans occurred in only one *erectus* population (most likely the African one) and that this fully modern form went on to subsume or replace remaining *erectus* stragglers. This latter pattern of events is described as the 'recent African origin' hypothesis (Stringer & Bräuer, 1994).

The limited fossil evidence available is apparently unable to distinguish between these competing hypotheses (see Waddle, 1994, however, for an example of a quantitative approach which may eventually resolve the palaeontological debate). The relative likelihood of either hypothesis characterising the actual course of human evolution can, however, be assessed according to the expected impact on genetic variation (in this instance, ß-globin sequence diversity) in present-day human populations. Under either scenario, for instance, ß-globin sequence variation should reflect an African origin. If the multiregional hypothesis is correct, however, long-standing genetic differences between major geographic regions (i.e. Asia and Africa) are expected. This is because drift acting in, even partially, subdivided populations (in order for regional populations to have remained part of the same species, some amount of gene flow must have occurred) will, over long periods of time, result in considerable genetic changes. A recent African origin of modern humans, on the other hand, would suggest that comparatively fewer ß-globin sequence differences should exist among human populations.

Whether modern human origins were ancient or recent, the nature of the species'

subsequent dispersal is also expected to be reflected in patterns of genetic similarities and differences among non-African populations. Although there are few explicit palaeontological predictions about the movement of anatomically modern humans following a recent African origin, several alternative scenarios have been proffered on the basis of genetic (primarily mitochondrial, but see also Nei & Roychoudhury, 1993) evidence. The 'weak Garden of Eden' hypothesis (Harpending et al., 1993) suggests that following the origin of humans in Africa 100,000 to 150,000 years ago, *H. s. sapiens* separated into two or more daughter populations, which existed as small, and partially isolated groups for tens of thousands of years prior to their simultaneous (and dramatic) increase in population size approximately 50,000 years ago. This hypothesis is contrasted with an alternative model, the 'strong Garden of Eden' theory, which posits a single Old World expansion of African *H. s. sapiens* shortly after the species originated.

The Geographic Origin of Human ß-globin Variation

Each of the hypotheses of human evolutionary origins discussed above posits an ultimate African origin of present-day human populations. In the multiregional hypothesis, *Homo erectus* evolves in Africa and disperses to the Old World, whereas recent African origin models, whether strong or weak Garden of Eden, have anatomically modern human beings evolving in the African *H. erectus* population and then expanding to other geographic regions. All three scenarios require the migration of descendant populations out of the African continent. In the course of such migrations, it is likely that some of the variation present in the African source population would have been lost, either as a consequence of founder effects (i.e. the migration of non-representative subsamples of the African gene pool) or population bottlenecks (i.e. the loss of variation through increased genetic drift in smaller migrant populations). Therefore, as long as the non-migrant source population remained roughly the same size, today there should be more, and older, genetic variation in Africa than in other human populations. Allele sequence diversity at the human ß-globin locus is, in fact, consistent with these predictions.

The level of ß-globin polymorphism in the Gambian population, measured in terms of the number of polymorphic sites relative to the total size of the sample (θ, or nucleotide diversity), was higher than in any of the other four populations investigated. As the influence of natural selection on nucleotide diversity at the ß-globin locus is negligible (Chapter 4), this observation suggests that drift has had a smaller impact relative to mutation in the Gambian population, i.e. that the African sample derives from a population which has been larger for longer. The converse situation, namely that recent reductions in population size (such as might have accompanied the founding of island populations) are responsible for the lower level of diversity in the non-African populations, is unlikely because estimates of nucleotide diversity in different non-African samples were the same.* Similarly, observed differences are probably not an artefact of the choice of populations investigated, because other surveys of human genetic variation have also found the greatest genetic diversity in African populations (Wainscoat et al., 1986; Cann, Stoneking, & Wilson, 1987; Stoneking et al., 1990; Merriwether et al., 1991; Horai & Hayasaka, 1990; Vigilant et al., 1989 & 1991; Bowcock et al., 1994; Relethford & Harpending, 1994; Martinson et al., in preparation).

Additional evidence for an African origin of ß-globin diversity derives from the geographical distribution of B-type ß-globin sequence haplotypes. Of the four major kinds of ß-globin sequence haplotypes identified, B-type alleles share the greatest number of similarities with the chimpanzee ß-globin gene (Savatier et al., 1985, J. Bond, personal communication). B-type sequence haplotypes, in concordance with neutral theory predictions that the oldest allele will be most frequent, were also the most common type of sequence haplotype observed in the populations surveyed, and the only sequence haplotype observed in all five populations was of type B. It is likely, therefore, that all present-day human ß-globin alleles derive from an ancestral, B-type, gene. In this context, it is interesting to note that two of the three B sequence haplotype subgroups, including the one most closely related to the chimpanzee ß-globin gene (i.e. sequence haplotypes 5B16 to 5B22, Figure 3.3) were confined

---

* Theoretical findings also suggest that the number of polymorphic sites in a population will not be substantially affected by short-term reductions in population size (Tajima, 1989b).

exclusively to the Gambian population. The African-specific distribution of what is likely to be the oldest ß-globin sequence lineage strongly supports an ultimate African origin for the observed variation.

Further sampling of ß-globin allelic sequence diversity, in both African and non-African populations, is required to confirm these preliminary findings. Of particular interest in this regard will be whether the 'ancestral' ß-globin lineage is observed in any non-African populations. Additional information, about the extent to which variation in all four major ß-globin sequence haplotype groups (allelic lineages which probably diverged prior to speciation, see below) is or is not highest for samples from the African continent, is also required. It will also be interesting to compare genetic differences among African populations with those found among populations in other major geographic regions. Inter-African genetic distances, estimated on the basis of α-globin RFLP haplotype variation, have been shown to be considerably larger than those observed among non-African (i.e. European, Asian, and Oceanic) populations (A. J. Boyce, personal communication).

The Time of Divergence of Modern Human Populations

ß-globin sequence diversity, in common with variation observed at other human loci, suggests an African origin for modern human populations. This finding is consistent both with the African origin of *Homo erectus* described by the multiregional hypothesis, and with the evolution of *H. s. sapiens* from an African population of *H. erectus*, or recent African origin model. What distinguishes these major theories of human evolutionary origins is, of course, their predictions regarding the time that human populations have been diverging since their common African origin. When ancestrally-related populations are separated from one another by geographic, cultural, or linguistic, barriers, they can quickly (depending on the size of the populations, the extent of their reproductive isolation, and the rate of mutation) accumulate regionally-specific polymorphisms, differences which reflect the time that has passed since the populations diverged. The distribution of ß-globin sequence variation within and

between sampled locales may therefore be used to infer the timing of the major geographic dispersals which followed human African origin.

A comparison of observed variation with interspecific (human-chimpanzee) sequence divergence suggests that polymorphic differences at the human ß-globin locus reflect mutations which have persisted for a considerable period of time. 1.4% of effectively silent sites differ between the GenBank (Bilofsky & Burks, 1988) human ß-globin consensus sequence (Poncz et al., 1983), a type A ß-globin gene, and an equivalent genomic region in the common chimpanzee species, *Pan troglodytes* (Savatier et al., 1985 & J. Bond, personal communication). As humans and chimpanzees are likely to have begun diverging approximately 5 million years ago (an estimate based on molecular data* from a large number of sources, Saitou & Ueda, 1994; also Horai et al., 1992), these differences suggest a rate of sequence divergence at the locus (disregarding interspecific length differences) of 0.28% per million years. Twenty-six polymorphic nucleotide sites (not insertion/deletion polymorphisms) were observed among the human ß-globin alleles investigated here, 4.7 of which varied between any two randomly chosen alleles, corresponding to an average sequence divergence of 0.18% and, therefore, an average age of polymorphism at the locus of 640,000 years.

The total time that ß-globin alleles have been diverging is likely to be much longer, however, as the two most divergent sequence haplotypes, 5A1 and 1D3, differ at 13 or 0.49% of sites. The total coalescence time, T, of ß-globin sequence variation (i.e. the length of time that human ß-globin alleles have been diverging from a single common ancestor), can be estimated as (Tajima, 1983, as presented by Templeton, 1993):

$$T = \frac{\theta(1+k)}{2\mu(1+n\theta)} \qquad (6.1)$$

with a variance of

$$\sigma^2 = \frac{\theta^2(1+k)}{4\mu^2(1+n\theta)^2} \qquad (6.1a)$$

---

* It has been argued that the palaeontological evidence is consistent with a slightly earlier human-chimpanzee divergence of 6 million years (Hill & Ward, 1988). The fragmentary nature of the fossil record, as well as uncertainty regarding the precise relationship between morphological and molecular evolution, make the use of the molecular-based divergence estimate preferable in this instance.

where θ is the average nucleotide diversity (in this instance, 0.18%), k is the total number of sequence differences between the two most divergent sequence haplotypes (13), 2μ is two-times the rate of mutation at the locus or the rate of sequence divergence (0.28% per million years), and n is the total number of effectively silent sites surveyed (2658). Using these figures, the total coalescence time for observed ß-globin variation is 1,580,000 ± 420,000 years. If we disregard polymorphisms in the first 500 bp of the sequenced region (which may be affected in some way by the high rate of interallelic recombination in the 5' flanking DNA), average sequence divergence is 0.13%, maximum pairwise divergence (k) is 10, and the total coalescence time is estimated as 1,350,000 ± 400,000 years.

Such ancient variation at the human ß-globin locus might, on initial inspection, appear consistent with the older origin of *H. s. sapiens* postulated by the multiregional hypothesis. However, unless the human species underwent a profound reduction in population size just as it was founded, so that all (or nearly all) of the variation inherited from the ancestral population was lost, there is no reason to assume that allelic divergence began when the species originated. Available evidence is, in fact, inconsistent with such an evolutionary scenario: polymorphic variation at major histocompatibility (MHC) loci, which has persisted for tens of millions of years, suggests that the effective population size of ancestral human populations was never smaller than 1000 individuals (Takahata, 1993; Ayala et al., 1994). Therefore, the age of ß-globin polymorphism, which may include variation which arose in an earlier, ancestral, species, is not likely to reflect the timing of the origin of anatomically modern humans.

Instead, the number of sequence differences specific to African and non-African populations, which represent genetic changes which occurred during, or after, geographic dispersal, provide a more reliable indication of the relative timing of modern human origins (the extent to which such diversity can be used to estimate the *actual* timing of speciation depends, of course, on how quickly after speciation the populations dispersed, as well as how much subsequent gene flow occurred between geographically-separated populations). Only 3 of the 15 silent polymorphisms observed

in the last 2500 bp of the sequenced region were confined solely to the non-African populations surveyed, another 3 were found only in the Gambian population. In order to derive the approximate amount of time consistent with such regionally-specific polymorphism, estimates of sequence coalescence, based on total (shared plus specific) variation in each major geographic region were compared to equivalent estimates calculated on the basis of common polymorphism alone. For example, the maximum pairwise divergence (*k*) among sequence haplotypes sampled from the four Asian populations is 10, resulting in an estimated coalescence time of 1,310,000 years. When the three Asian-specific polymorphisms are disregarded, however, *k* for the same sample of alleles is only 8, giving the proportionally smaller coalescence time of 1,070,000 years. This suggests that the Asian-specific sequence differences have arisen over the course of approximately 240,000 years (the difference between the two coalescence estimates). Similar calculations for the Gambian variation suggest that observed African-specific polymorphisms are only 125,000 years old.

The difference between these regional estimates may derive from the more limited sampling of African variation represented by the Gambian sample, or could simply reflect the wide range of values associated with the coalescence estimates. Regardless of the discrepancy, either estimate is far too recent to be consistent with the 1 to 2 million year differentiation predicted by the multiregional hypothesis. The only way the observed pattern of ß-globin sequence variation could be reconciled with the multiregional model would be if there had been considerable gene flow among geographically-dispersed African and Asian populations. Such gene flow, however, would be expected to result in the random dispersal of mutations arising in different geographic regions so that, for instance, some Asian populations would share polymorphisms with Africa while other populations would not. Instead, as discussed in Chapter 5, the sites found to be shared by African and non-African populations in this study were ubiquitous (8 of the 9 common polymorphisms in the 2500 bp region were observed in all five populations surveyed), suggesting that the observed similarities are a consequence of recent divergence from a single source population, rather than long-distance gene flow.

The observed distribution of ß-globin nucleotide polymorphism is therefore consistent with a recent common African origin, confirming results presented by other investigators, which provide little or no support for the multiregional hypothesis and which, collectively, suggest an age of divergence of anatomically modern humans of between 100,000 and 200,000 years ago (Cann, Stoneking, & Wilson, 1987; Horai & Hayasaka, 1990; Vigilant et al., 1989 & 1991; Nei & Roychoudhury, 1993; Ruvolo et al., 1993; Mountain & Cavalli-Sforza, 1994). The current survey, of course, is limited in that it provides no information about the nature of genetic similarities and differences among Asian, or African, and European populations. Equivalent investigation of ß-globin sequence polymorphism in a British population has recently been completed (J. A. Schneider, personal communication) and initial results suggest that British and Southeast Asian ß-globin alleles are, in fact, closely related, particularly at the nucleotide level. Thus, it appears unlikely that additional data will significantly modify the conclusions reported here (further sampling, however, will lend greater confidence to inferred estimates of African/non-African divergence times).

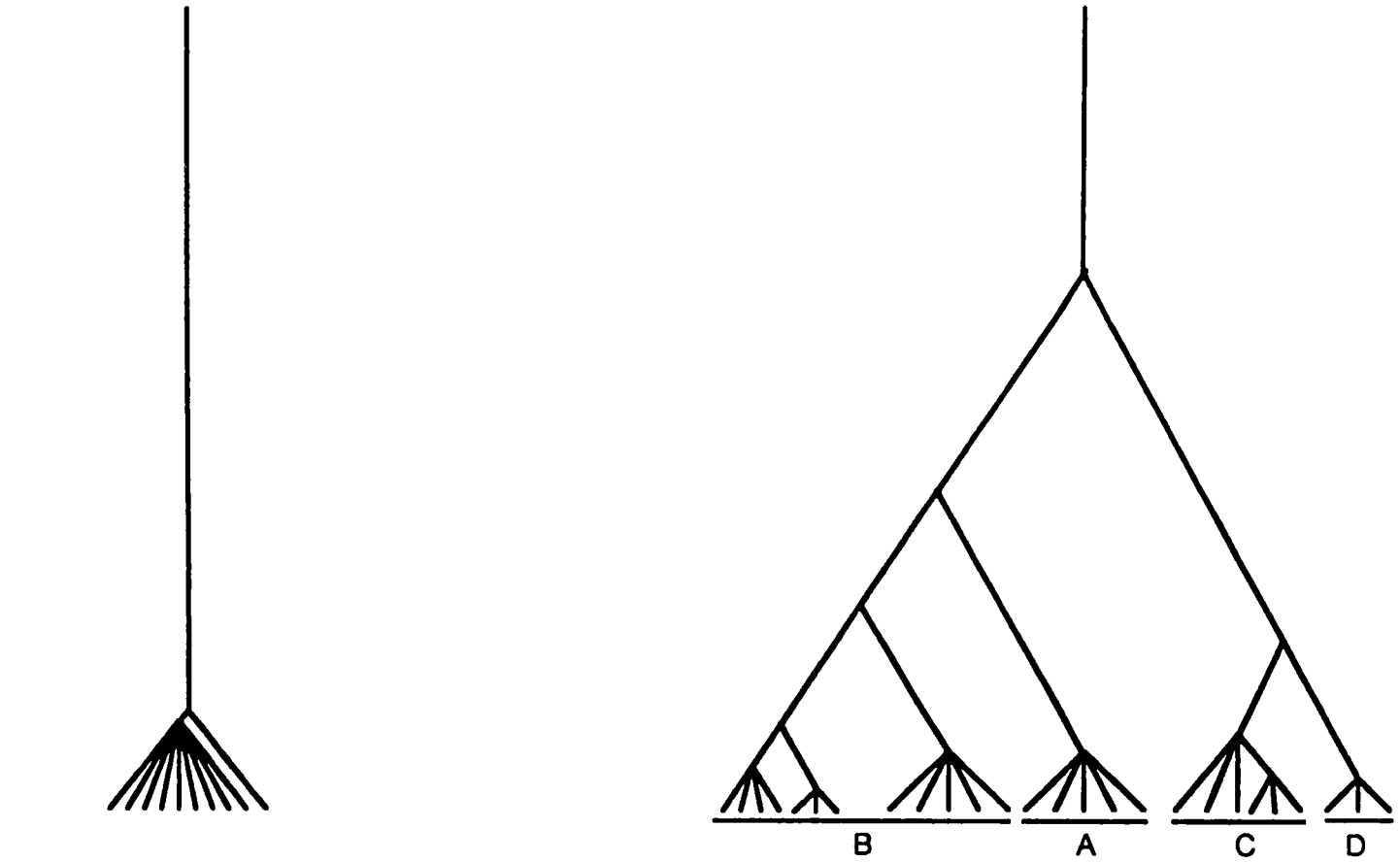
Evidence for Population Expansions

DNA sequence variation at the human ß-globin locus is clearly most consistent with the recent African origin hypothesis of human evolutionary origins. As discussed earlier in this chapter, a further evolutionary theory, which attempts to describe the events that followed speciation, has been suggested on the basis of observed mtDNA sequence diversity, the weak Garden of Eden hypothesis (Harpending et al., 1993). This theory suggests that early human populations split into several small, geographically-isolated, groups shortly after speciation and slowly differentiated over tens of thousands of years. Then, approximately 50,000 years ago these populations underwent simultaneous, and apparently extremely rapid, increases in population size (accompanied by geographic expansion), due, it is suggested, to some cultural advantage, i.e. the archaeological record shows that a shift in tool use began 40,000 to 50,000 years ago (Klein, 1992). A lack of corroborative palaeontological evidence, however, combined with questions regarding the reliability of statistical inferences drawn from the mtDNA data (Marjoram

& Donnelly, 1994), mean that the theory is far from widely accepted. The relationship of ß-globin sequence diversity to the hypothesis' predictions is therefore of considerable evolutionary interest.

The weak Garden of Eden hypothesis bases its inferences on the expected effect of population expansion on genetic diversity in populations, namely that increases in effective population size result in an increased likelihood of retaining polymorphic variation because the effects of random genetic drift are proportionately diminished. Therefore, with population expansion, a single sequence lineage (or type of allele) diverges rapidly, resulting in what has been described as a 'star-like phylogeny' (Di Rienzo & Wilson, 1991) (Figure 6.1). All human populations show such patterns of variation when mtDNA sequence diversity is considered (Rogers & Jorde, 1994), and it is the number of sequence differences between divergent mtDNA lineages which have been used to infer the timing of expansion events (Harpending et al., 1993). As Figure 6.1 also illustrates, however, a different pattern of allelic relationship characterises variation at the human ß-globin locus: ß-globin alleles fall into four major groups, whose sequence similarities and differences suggest a bifurcating, rather than star-shaped, gene tree.

Differences in patterns of mtDNA and ß-globin allelic sequence divergence may relate to the timing of putative population expansion events relative to the overall persistence of allelic variation at either loci. The mitochondrial DNA genome, for instance, because it is haploid and maternally-inherited, is much more susceptible to the effects of genetic drift than nuclear DNA. As a consequence, polymorphic differences persist for shorter periods and observed sequence lineages coalesce to a single common ancestor sooner (the coalescence time of mtDNA is approximately 250,000 years, assuming a 5 million year old chimp-human ancestor, Ruvolo et al., 1993). The star-shaped phylogeny observed for mtDNA types therefore reflects the 'early' impact of population expansion on presently-detectable sequence differences. At the ß-globin gene, however, major sequence lineages have been diverging for a much longer period (as discussed above, approximately 1.3 million years), so expansions in the last 50,000 years will have only affected the last 5% of the tree (i.e. the tips of the



Mitochondrial DNA **ß-Globin**

Figure 6.1 Inferred phylogenetic relationships among alleles identified at the human mitochondrial DNA locus (based on a similar diagram presented by Rogers & Jorde, 1994), as well as in the last 2500 bp the of the 3 kb ß-globin gene region (a preliminary interpretation of sequence relationships identified using the tree building computer program PAUP, or Phylogenetic Analysis with Parsimony [Swofford, 1991]). The tree upon which the ß-globin schematic is based is given in Appendix C.

branches).

It is impossible, at present, to determine whether observed allelic sequence differences at the ß-globin locus are, in fact, consistent with major increases in population size in the last 50,000 to 100,000 years. Although the considerable within-lineage (or, sequence haplotype group) diversity identified could, in part, reflect the effects of population expansion, most of this variation derives from the impact of interallelic recombination on the ß-globin gene, whose effects are difficult to interpret phylogenetically. The significance of lineage-specific nucleotide polymorphisms, i.e. sites found to vary in only a single haplotype group, is also unclear. Many of the observed lineage-specific differences are rare, which might be consistent with their occurrence in recently expanded human populations. However, two-thirds of these rare changes (and, indeed, the only rare differences observed outside the Gambia) are found in the first 550 bp of the ß-globin gene region, so that the possibility that their occurrence is associated with the recombinatorial activity in the 5' flanking DNA, rather than with particular demographic effects, cannot be discounted. A better understanding of the persistence of polymorphism in recombinogenic genomic regions, and the relationship of this persistence to changes in population size, is required.

In contrast, the global distribution of major sequence lineages, particularly C- and D-type sequence haplotypes, may be consistent with the impact of population expansion. As revealed by the sequence analyses conducted here, as well as earlier gene framework studies, C-type ß-globin alleles are much more common than D-type genes in Asian populations, whereas more D-type sequences and fewer Cs are observed in European groups.* The widespread geographic distribution of regionally-specific sequence haplotype patterns suggests that the observed genetic differences originated in founding Asian and European populations. In order for genetic drift to have resulted in their differentiation, these early populations must have been genetically-isolated, and perhaps relatively small in size, so that by chance the distribution of the (judging by

---

* The latter pattern, which is suggested on the basis of variation observed in normal and thalassaemic Mediterranean ß-globin genes (Orkin et al., 1982), is supported by preliminary analysis of ß-globin sequence variation in a British population, which, unlike equivalent samples from Southeast Asia, has been found to possess more D- than C-type sequence haplotypes (J. A. Schneider, personal communication).

the composition of the Gambian population) least common types of sequence haplotypes in each group changed. Subsequent population expansion (occurring prior to, or in conjunction with, continental dispersal), which would have the effect of 'freezing' the relative proportions of C- and D-type ß-globin alleles in all descendant populations, is then the most plausible explanation for the observed/reported Asian and European sequence haplotype distributions.

Available evidence therefore suggests that variation at the ß-globin locus may be consistent with the expansion of modern humans from geographically-restricted founding populations. These tentative inferences must, of course, be confirmed by additional investigation of variation in other human populations. An indirect method of distinguishing major types of ß-globin sequence haplotypes, which would permit the examination of large numbers of samples in many geographic regions, would be sufficient to explore the extent to which patterns of allelic variation found on different continents are, in fact, homogeneous. In this regard, current disagreement among investigators as to whether patterns of mtDNA diversity in African populations reflect the effects of population expansion (compare Di Rienzo & Wilson, 1991 and Harpending et al., 1993), make the analysis of ß-globin variation in African populations particularly relevant. Finally, more fine-scale studies of variation in different geographic regions, combined with theoretical exploration of the expected nature of nuclear DNA variation in constant-sized and expanding populations, will be required to determine the extent to which observed ß-globin sequence differences are consistent with recent increases in population size.

ß-Globin Sequence Variation and Modern Human Origins

As the preceding discussion has illustrated, allelic sequence diversity at the human ß-globin locus is consistent, in most respects, with the weak Garden of Eden hypothesis. The largest amount of nucleotide diversity was observed in the Gambian sample, suggesting an African origin for present-day ß-globin variation, and the sequence differences observed between African and non-African populations were found to be

compatible with the recent (i.e. less than 250,000 years) divergence of modern human populations. In addition, although similarities and differences among ß-globin sequence haplotypes were impossible to interpret in terms of the effects of population expansion, the geographical distribution of C- and D- type ß-globin alleles did appear to reflect the expansion of human populations from semi-isolated founding groups. As such, these conclusions coincide with similar inferences drawn on the basis of other DNA data,* including ß-globin (Wainscoat et al., 1986; Long et al., 1990) and α-globin RFLP haplotypes (Martinson et al., in preparation), unlinked RFLP sites (Bowcock, et al., 1991b; Mountain & Cavalli-Sforza, 1994) and CA-repeat polymorphisms (Bowcock et al., 1994), as well as both RFLP haplotype (Cann, Stoneking, & Wilson, 1987; Merriwether et al., 1991) and DNA sequence variation (Horai & Hayasaka, 1990; Vigilant et al., 1989; 1991; Ruvolo et al., 1993; Harpending et al., 1993) at the human mtDNA locus.

The fact that ß-globin sequence variation leads to the same conclusions as compatible sequence-level analysis of mtDNA diversity is important because current interpretations of mtDNA data have been questioned on both palaeontological (Thorne & Wolpoff, 1992; Frayer et al., 1994) and analytical (Excoffier, 1990; Maddison, 1991; Templeton, 1993; Marjoram & Donnelly, 1994) grounds. It is difficult, however, to explain the concordance of results obtained from the two loci, which differ markedly in mutation rate, the effects of recombination, and susceptibility to drift, in anything other than demographic terms. In this context, it is clear that inferences based on variation observed in either genomic region have been greatly strengthened by having equivalent data from an unlinked locus for comparison. Mitochondrial DNA, for instance, although well-suited (by virtue of its high mutation rate) to the investigation of recent genetic changes in human populations, is also so susceptible to the effects of random genetic drift that present-day variation offers only a limited observational window on the time period likely to be associated with human evolutionary origins. ß-globin diversity, in contrast, possesses a much greater time-depth than mtDNA (thereby lending itself to more reliable estimates of the timing of human population divergence),

---

* Results are also consistent with a recently published analysis of craniometric (genetic) variation in human populations (Relethford & Harpending, 1994).

but is not variable enough to be amenable to the analysis of recent evolutionary changes.

The investigation of human origins will benefit from similar, sequence-level, analysis of variation at other unlinked nuclear loci. Besides confirming, or otherwise refining, current evolutionary interpretations, additional data should also help clarify the effects of both selection and recombination on human nuclear DNA variation. Such clarification is important because, although it appears likely that the human ß-globin locus is evolving neutrally, the impact of selection on polymorphic variation at the locus cannot be ruled out altogether. Similarly, it is unclear to what extent the observed localised influence of interallelic recombination at the ß-globin gene, which is unlikely to be typical of the majority of human nuclear loci, is affecting the overall level and/or pattern of variation at the locus. Sequence-level analysis of variation in genomic regions subject to differing levels of functional constraint (such as pseudogenes) and selective effects (the R-T locus 5' of the human δ-globin gene [Maeda, Bliska, & Smithies, 1983] is a nearby example), as well as found in different recombinatorial environments (for instance, the α-globin gene cluster, which is currently under investigation by J. Norwich in our laboratory) will go some way toward elucidating the impact of diverse evolutionary processes on human genetic variation.

The extensive nucleotide and sequence haplotype variation identified in the course of this research amply demonstrate the power of DNA sequence analysis for human population genetic study. With the routine detection of allelic sequence diversity now a reality, and with the concomitant (and on-going) development of new theoretical and statistical tools for analysing such variation, there is much to encourage future investigation. With luck, the analyses initiated here will be extended to the study of variation at many other nuclear loci, furthering the aims of human evolutionary analysis, and continuing the deeply engaging process of self-discovery which has so-characterised human population genetic investigation.

APPENDICES

Appendix A

Standard Methods and Materials

A number of methodological approaches to the investigation of allelic sequence variation at the human ß-globin gene were explored before a standard protocol was adopted. Here, the method which ultimately proved successful is outlined, in detail. All of the data presented in this thesis (except where otherwise stated) derive from the application of these methods.

Overview of Methodology

Direct sequence analysis of Polymerase Chain Reaction (PCR)-amplified genomic DNA was used to investigate the nucleotide sequence of allelic copies of the human ß-globin gene in diploid samples.

DNA for sequencing was amplified using PCR primers specific to the ß-globin region, one of which was biotinylated. The resulting biotinylated product was bound to streptavidin-coated magnetic beads and subsequently purified to single-stranded template. This diploid template was sequenced by the dideoxy chain termination method using multiple internal sequencing primers. Compound heterozygotes were re-amplified using allele-specific PCR amplification primers. The resulting allele-specific product was sequenced to directly determine the sequence haplotype of one allele. The sequence haplotype of the other allele was then derived by comparison with the original diploid sequence, where possible. In other instances the second allele was specifically amplified and sequenced.

In addition, the status of flanking restriction fragment length polymorphisms (RFLPs) was

investigated by PCR amplification and digestion of individual sites.

Amplification of Genomic Template

Whole genomic DNA, isolated from blood, was PCR-amplified to obtain quantities of DNA sufficient for sequencing.

Extraction of Genomic DNA from Blood

Genomic DNAs used in this study were extracted either from umbilical cord, or venous, human blood samples, using standard Proteinase K digestion/phenol extraction chemistry (Old and Higgs, 1983):

Each 10 ml blood sample (fresh or from -20 °C) was haemolysed in 50 ml H_20. The lysed erythrocytes were centrifuged (3000 rpm, 15 min, 4 °C) and the haemoglobin-containing supernatant discarded. The pelleted cells were resuspended in 25 ml chilled Triton/sucrose 'Pre-lysis' Buffer (0.32 M sucrose, 10 mM Tris-Cl pH 7.6, 5 mM $MgCl_2$, 0.1% Triton X-100) to lyse leukocyte cell membranes. Centrifugation (10 min) pelleted the cell nuclei, the supernatant was removed, and the nuclear pellet washed with 5 ml 0.1% Phosphate Buffered Saline (PBS) (0.13 M NaCl, 5 mM KCl, and 7.4 mM $MgCl_2$-$6H_20$). Following centrifugation (2000 rpm, 5 min, 4 °C), the washed cell nuclei were lysed by incubation in 10 ml Lysis Buffer (10 mM Tris-borate pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5% Sodium dodecyl sulphate (SDS), 0.5 mg/ml Proteinase K) at 37 °C overnight.

The lysed nuclei were extracted once with phenol:water:chloroform (70:20:10 v/v, Rathburn Chemicals Ltd.) and twice with chloroform, to remove peptides, lipids, and other contaminates. Finally, the DNA was precipitated with 0.1 vol (vol) 3 M NaOAc pH 5.8 and 2 vols ice-cold 100% ethanol, spun (3500 rpm, 30 min, 4 °C), washed with 1 vol 70% ethanol, spun (15 min), air-dried, and resuspended in 500 µl sterile TE pH 7.6. The resulting genomic DNA (typically 200-500 µg) was stored at -20 °C.

The majority of samples analysed here were collected and extracted by the supplying investigator, for the purpose of other research (see 'Samples Investigated' section).

Polymerase Chain Reaction Amplification (Diploid Template)

A 3.067 kilobase (kb) fragment (Figure A.1), containing all of the human ß-globin gene and 1.4 kb of flanking DNA, was amplified from each genomic DNA sample using the Polymerase Chain Reaction or PCR (Saiki, et al. 1985).

A single pair of oligonucleotide primers was used to amplify the 3 kb segment. They are designated 5' AP (Amplification Primer) and 3' AP in Figure A.1, and anneal respectively to DNA ~850 basepairs (bp) 5' of the human ß-globin gene cap site, and ~600 bp 3' of the gene's polyadenylation signal. The primer sequences correspond to positions 10547-10572 and 13592-13613 of the ß-globin sequence reported by Poncz, et al. (1983).

500-1000 ng genomic DNA per 100 µl reaction was amplified using either 5' AP and 3' B'd (Biotinylated) AP (4 reactions) or 5' B'd AP and 3' AP (1 reaction). One of the two amplification primers was biotinylated (see 'Oligonucleotides Used in Study' section) to 'label' the PCR product for subsequent manipulation. 10 pmol of each primer, 1.85 µg gene 32 protein ('gp32', Pharmacia LKB Biotechnology), and 2 units of *Taq* polymerase (Advanced Biotechnologies Ltd.) were used per 100 µl reaction (50 mM KCl, 10 mM Tris-Cl pH 8.4, 1.5 mM $MgCl_2$, 0.001% gelatin [Sigma Chemical Co., Ltd.], 200 mM dNTPs [Pharmacia LKB Biotechnology]). Each reaction was topped with 50 µl mineral oil (Sigma Chemical Co., Ltd.) to prevent evaporation.

The majority of reactions were performed in a DNA Thermal Cycler (Perkin Elmer Cetus); others in a Omnigene Temperature Cycler (Hybaid Ltd.). PCR reaction conditions, in either machine, were as follows: an initial cycle of 4 min at 94 °C (to thoroughly denature the genomic template), 2 min at 60 °C (for primer annealing), and 5 min at 72 °C (the optimal temperature for *Taq* polymerase extension), then 40 cycles

```
                    20                  40                  60                  80
    *     * 5' AP *     *  5' SP 1    *     *     *     *     *     *     *     *
AGTCCAGGCAGAAACAGTTA GATGTCCCCAGTTAACCTCC TATTTGACACCACTGATTAC CCCATTGATAGTCACACTTT
TCAGGTCCGTCTTTGTCAAT CTACAGGGGTCAATTGGAGG ATAAACTGTGGTGACTAATG GGGTAACTATCAGTGTGAAA

                   100                 120                 140                 160
    *     *     *     *     *     *     *     *     *     *     *     *     *     *
GGGTTGTAAGTGACTTTTTA TTTATTTGTATTTTTGACTG CATTAAGAGGTCTCTAGTTT TTTATCTCTTGTTTCCCAAA
CCCAACATTCACTGAAAAAT AAATAAACATAAAAACTGAC GTAATTCTCCAGAGATCAAA AAATAGAGAACAAAGGGTTT

                   180                 200                 220                 240
    *     *     * 5' SP 1a    *     *     *     *     *     *     *     *     *     *
ACCTAATAAGTAACTAATGC ACAGAGCACATTGATTTGTA TTTATTCTATTTTTAGACAT AATTTATTAGCATGCATGAG
TGGATTATTCATTGATTACG TGTCTCGTGTAACTAAACAT AAATAAGATAAAAATCTGTA TTAAATAATCGTACGTACTC

                   260                 280                 300                 320
    *     *     *     *     *     *     *     *     *     *     *     *     *     *
CAAATTAAGAAAAACAACAA CAAATGAATGCATATATATG TATATGTATGTGTGTATATA TACACATATATATATATATT
GTTTAATTCTTTTTGTTGTT GTTTACTTACGTATATATAC ATATACATACACACATATAT ATGTGTATATATATATATAA

                   340                 360                 380                 400
    *     *     *     *     *     *     *     *     *     * 5' SP 2 *     *     *
TTTTTTCTTTTCTTACCAGA AGGTTTTAATCCAAATAAGG AGAAGATATGCTTAGAACTG AGGTAGAGTTTTCATCCATT
AAAAAAGAAAAGAATGGTCT TCCAAAATTAGGTTTATTCC TCTTCTATACGAATCTTGAC TCCATCTCAAAAGTAGGTAA

                   420                 440                 460                 480
    *     *     *     *     *     *     *     *     *     *     *     *     *     *
CTGTCCTGTAAGTATTTTGC ATATTCTGGAGACGCAGGAA GAGATCCATCTACATATCCC AAAGCTGAATTATGGTAGAC
GACAGGACATTCATAAAACG TATAAGACCTCTGCGTCCTT CTCTAGGTAGATGTATAGGG TTTCGACTTAATACCATCTG

                   500                 520                 540                 560
    *     *     *     *     *     *     *     *     *     *     *     *     *     *
AAAACTCTTCCACTTTTAGT GCATCAATTTCTTATTTGTG TAATAAGAAAATTGGGAAAA CGATCTTCAATATGCTTACC
TTTTGAGAAGGTGAAAATCA CGTAGTTAAAGAATAAACAC ATTATTCTTTTAACCCTTTT GCTAGAAGTTATACGAATGG

                   580                 600                 620                 640
    *     *     *     *     *     *     *     *     *     *     *     *     *     *
AAGCTGTGATTCCAAATATT ACGTAAATACACTTGCAAAG GAGGATGTTTTTAGTAGCAA TTTGTACTGATGGTATGGGG
TTCGACACTAAGGTTTATAA TGCATTTATGTGAACGTTTC CTCCTACAAAAATCATCGTT AAACATGACTACCATACCCC
                                                                      3' SP 8
                   660                 680                 700                 720
    *     *     *     *     *     *     *     * 5' SP 3     *     *     *     *     *
CCAAGAGATATATCTTAGAG GGAGGGCTGAGGGTTTGAAG TCCAACTCCTAAGCCAGTGC CAGAAGAGCCAAGGACAGGT
GGTTCTCTATATAGAATCTC CCTCCCGACTCCCAAACTTC AGGTTGAGGATTCGGTCACG GTCTTCTCGGTTCCTGTCCA

                   740                 760                 780                 800
    *     *     *     *     *     *     *     *     *     *     *     *     *     *
ACGGCTGTCATCACTTAGAC CTCACCCTGTGGAGCCACAC CCTAGGGTTGGCCAATCTAC TCCCAGGAGCAGGGAGGGCA
TGCCGACAGTAGTGAATCTG GAGTGGGACACCTCGGTGTG GGATCCCAACCGGTTAGATG AGGGTCCTCGTCCCTCCCGT

                   820                 840                 860                 880
    *     *     *     *     *     *     *     *     *     *     *     *     *     *
GGAGCCAGGGCTGGGCATAA AAGTCAGGGCAGAGCCATCT ATTGCTTACATTTGCTTCTG ACACAACTGTGTTCACTAGC
CCTCGGTCCCGACCCGTATT TTCAGTCCCGTCTCGGTAGA TAACGAATGTAAACGAAGAC TGTGTTGACACAAGTGATCG

                   900                 920                 940                 960
    *     *     *     *     *     *     *     *     *     *     *     *     *     *
AACCTCAAACAGACACCATG GTGCACCTGACTCCTGAGGA GAAGTCTGCCGTTACTGCCC TGTGGGGCAAGGTGAACGTG
TTGGAGTTTGTCTGTGGTAC CACGTGGACTGAGGACTCCT CTTCAGACGGCAATGACGGG ACACCCCGTTCCACTTGCAC
      3' SP 7
                   980                1000                1020                1040
    *     *     *     *     *     *     *     *     *     *     *     *     *     *
GATGAAGTTGGTGGTGAGGC CCTGGGCAGGTTGGTATCAA GGTTACAAGACAGGTTTAAG GAGACCAATAGAAACTGGGC
CTACTTCAACCACCACTCCG GGACCCGTCCAACCATAGTT CCAATGTTCTGTCCAAATTC CTCTGGTTATCTTTGACCCG
```

Figure A.1 DNA sequence of the 3.067 kb PCR fragment, showing the positions of the 5' and 3' sequencing primers.

```
                1060                1080                1100                1120
    *    *    *    *  5' SP 4    *    *    *    *    *    *    *    *    *    *
ATGTGGAGACAGAGAAGACT CTTGGGTTTCTGATAGGCAC TGACTCTCTCTGCCTATTGG TCTATTTTCCCACCCTTAGG
TACACCTCTGTCTCTTCTGA GAACCCAAAGACTATCCGTG ACTGAGAGAGACGGATAACC AGATAAAAGGGTGGGAATCC

                1140                1160                1180                1200
    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *
CTGCTGGTGGTCTACCCTTG GACCCAGAGGTTCTTTGAGT CCTTTGGGGATCTGTCCACT CCTGATGCTGTTATGGGCAA
GACGACCACCAGATGGGAAC CTGGGTCTCCAAGAAACTCA GGAAACCCCTAGACAGGTGA GGACTACGACAATACCCGTT

                1220                1240                1260                1280
    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *
CCCTAAGGTGAAGGCTCATG GCAAGAAAGTGCTCGGTGCC TTTAGTGATGGCCTGGCTCA CCTGGACAACCTCAAGGGCA
GGGATTCCACTTCCGAGTAC CGTTCTTTCACGAGCCACGG AAATCACTACCGGACCGAGT GGACCTGTTGGAGTTCCCGT

                1300                1320                1340                1360
    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *
CCTTTGCCACACTGAGTGAG CTGCACTGTGACAAGCTGCA CGTGGATCCTGAGAACTTCA GGGTGAGTCTATGGGACCCT
GGAAACGGTGTGACTCACTC GACGTGACACTGTTCGACGT GCACCTAGGACTCTTGAAGT CCCACTCAGATACCCTGGGA
     3' SP 6
                1380                1400                1420                1440
    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *
TGATGTTTTCTTTCCCCTTC TTTTCTATGGTTAAGTTCAT GTCATAGGAAGGGGAGAAGT AACAGGGTACAGTTTAGAAT
ACTACAAAAGAAAGGGGAAG AAAAGATACCAATTCAAGTA CAGTATCCTTCCCCTCTTCA TTGTCCCATGTCAAATCTTA

                1460                1480                1500                1520
  5' SP 5   *    *    *    *    *    *    *    *    *    *    *    *    *    *    *
GGGAAACAGACGAATGATTG CATCAGTGTGGAAGTCTCAG GATCGTTTTAGTTTCTTTTA TTTGCTGTTCATAACAATTG
CCCTTTGTCTGCTTACTAAC GTAGTCACACCTTCAGAGTC CTAGCAAAATCAAAGAAAAT AAACGACAAGTATTGTTAAC

                1540                1560                1580                1600
    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *
TTTTCTTTTGTTTAATTCTT GCTTTCTTTTTTTTTCTTCT CCGCAATTTTTACTATTATA CTTAATGCCTTAACATTGTG
AAAAGAAAACAAATTAAGAA CGAAAGAAAAAAAAAGAAGA GGCGTTAAAAATGATAATAT GAATTACGGAATTGTAACAC

                1620                1640                1660                1680
    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *
TATAACAAAAGGAAATATCT CTGAGATACATTAAGTAACT TAAAAAAAAACTTTACACAG TCTGCCTAGTACATTACTAT
ATATTGTTTTCCTTTATAGA GACTCTATGTAATTCATTGA ATTTTTTTTTGAAATGTGTC AGACGGATCATGTAATGATA
                                                              3' SP 5
                1700                1720                1740                1760
    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *
TTGGAATATATGTGTGCTTA TTTGCATATTCATAATCTCC CTACTTTATTTTCTTTTATT TTTAATTGATACATAATCAT
AACCTTATATACACACGAAT AAACGTATAAGTATTAGAGG GATGAAATAAAAGAAAATAA AAATTAACTATGTATTAGTA

                1780                1800                1820                1840
    *    *    *    *    *    *    *    *    *    *    *    *  5' SP 6  *    *
TATACATATTTATGGGTTAA AGTGTAATGTTTTAATATGT GTACACATATTGACCAAATC AGGGTAATTTTGCATTTGTA
ATATGTATAAATACCCAATT TCACATTACAAAATTATACA CATGTGTATAACTGGTTTAG TCCCATTAAAACGTAAACAT

                1860                1880                1900                1920
    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *
ATTTTAAAAAATGCTTTCTT CTTTTAATATACTTTTTTGT TTATCTTATTTCTAATACTT TCCCTAATCTCTTTCTTTCA
TAAAATTTTTTACGAAAGAA GAAAATTATATGAAAAAACA AATAGAATAAAGATTATGAA AGGGATTAGAGAAAGAAAGT

                1940                1960                1980                2000
    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *
GGGCAATAATGATACAATGT ATCATGCCTCTTTGCACCAT TCTAAAGAATAACAGTGATA ATTTCTGGGTTAAGGCAATA
CCCGTTATTACTATGTTACA TAGTACGGAGAAACGTGGTA AGATTTCTTATTGTCACTAT TAAAGACCCAATTCCGTTAT
                                                              3' SP 4
                2020                2040                2060                2080
    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *    *
GCAATATTTCTGCATATAAA TATTTCTGCATATAAATTGT AACTGATGTAAGAGGTTTCA TATTGCTAATAGCAGCTACA
CGTTATAAAGACGTATATTT ATAAAGACGTATATTTAACA TTGACTACATTCTCCAAAGT ATAACGATTATCGTCGATGT
```

Figure A.1 continued

```
                2100                2120                2140                2160
    *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
ATCCAGCTACCATTCTGCTT TTATTTTATGGTTGGGATAA GGCTGGATTATTCTGAGTCC AAGCTAGGCCCTTTTGCTAA
TAGGTCGATGGTAAGACGAA AATAAAATACCAACCCTATT CCGACCTAATAAGACTCAGG TTCGATCCGGGAAAACGATT

                2180                2200                2220                2240
    *   *   *   *   *   *   *   *   *   *   * 5' SP 7 *   *   *   *
TCATGTTCATACCTCTTATC TTCCTCCCACAGCTCCTGGG CAACGTGCTGGTCTGTGTGC TGGCCCATCACTTTGGCAAA
AGTACAAGTATGGAGAATAG AAGGAGGGTGTCGAGGACCC GTTGCACGACCAGACACACG ACCGGGTAGTGAAACCGTTT

                2260                2280                2300                2320
    *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
GAATTCACCCCACCAGTGCA GGCTGCCTATCAGAAAGTGG TGGCTGGTGTGGCTAATGCC CTGGCCCACAAGTATCACTA
CTTAAGTGGGGTGGTCACGT CCGACGGATAGTCTTTCACC ACCGACCACACCGATTACGG GACCGGGTGTTCATAGTGAT

                2340                2360                2380                2400
    *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
AGCTCGCTTTCTTGCTGTCC AATTTCTATTAAAGGTTCCT TTGTTCCCTAAGTCCAACTA CTAAACTGGGGGATATTATG
TCGAGCGAAAGAACGACAGG TTAAAGATAATTTCCAAGGA AACAAGGGATTCAGGTTGAT GATTTGACCCCCTATAATAC
                                           3' SP 3
                2420                2440                2460                2480
    *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
AAGGGCCTTGAGCATCTGGA TTCTGCCTAATAAAAAACAT TTATTTTCATTGCAATGATG TATTTAAATTATTTCTGAAT
TTCCCGGAACTCGTAGACCT AAGACGGATTATTTTTTGTA AATAAAAGTAACGTTACTAC ATAAATTTAATAAAGACTTA

                2500                2520                2540                2560
    *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
ATTTTACTAAAAAGGGAATG TGGGAGGTCAGTGCATTTAA AACATAAAGAAATGAAGAGC TAGTTCAAACCTTGGGAAAA
TAAAATGATTTTTCCCTTAC ACCCTCCAGTCACGTAAATT TTGTATTTCTTTACTTCTCG ATCAAGTTTGGAACCCTTTT

                2580                2600                2620                2640
    *   *   *   *  5' SP 8  *   *   *   *   *   *   *   *   *   *   *   *
TACACTATATCTTAAACTCC ATGAAAGAAGGTGAGGCTGC AAACAGCTAATGCACATTGG CAACAGCCCTGATGCCTATG
ATGTGATATAGAATTTGAGG TACTTTCTTCCACTCCGACG TTTGTCGATTACGTGTAACC GTTGTCGGGACTACGGATAC
                                                                3' SP 2
                2660                2680                2700                2720
    *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
CCTTATTCATCCCTCAGAAA AGGATTCAAGTAGAGGCTTG ATTTGGAGGTTAAAGTTTTG CTATGCTGTATTTTACATTA
GGAATAAGTAGGGAGTCTTT TCCTAAGTTCATCTCCGAAC TAAACCTCCAATTTCAAAAC GATACGACATAAAATGTAAT

                2740                2760                2780                2800
    *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
CTTATTGTTTTAGCTGTCCT CATGAATGTCTTTTCACTAC CCATTTGCTTATCCTGCATC TCTCAGCCTTGACTCCACTC
GAATAACAAAATCGACAGGA GTACTTACAGAAAAGTGATG GGTAAACGAATAGGACGTAG AGAGTCGGAACTGAGGTGAG

                2820                2840                2860                2880
    *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
AGTTCTCTTGCTTAGAGATA CCACCTTTCCCCTGAAGTGT TCCTTCCATGTTTTACGGCG AGATGGTTTCTCCTCGCCTG
TCAAGAGAACGAATCTCTAT GGTGGAAAGGGGACTTCACA AGGAAGGTACAAAATGCCGC TCTACCAAAGAGGAGCGGAC

                2900                2920                2940                2960
    *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
GCCACTCAGCCTTAGTTGTC TCTGTTGTCTTATAGAGGTC TACTTGAAGAAGGAAAAACA GGGGGCATGGTTTGACTGTC
CGGTGAGTCGGAATCAACAG AGACAACAGAATATCTCCAG ATGAACTTCTTCCTTTTTGT CCCCCGTACCAAACTGACAG
                                                  3' SP 1
                2980                3000                3020                3040
    *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
CTGTGAGCCCTTCTTCCCTG CCTCCCCCACTCACAGTGAC CCGGAATCTGCAGTGCTAGT CTCCCGGAACTATCACTCTT
GACACTCGGGAAGAAGGGAC GGAGGGGGTGAGTGTCACTG GGCCTTAGACGTCACGATCA GAGGGCCTTGATAGTGAGAA

                3060
    *   *   *   *   *
TCACAGTCTGCTTTGGAAGG ACTGGGC
AGTGTCAGACGAAACCTTCC TGACCCG
          3' AP
```

Figure A.1 continued

of 1 min at 94 °C, 2 min at 60 °C, and 5 min at 72 °C, followed by a final cycle of 1 min at 94 °C, 2 min at 60 °C, and 10 min at 72 °C (a longer final extension helped to bring all primer extensions to full-length completion). Reactions were held either at 10 °C or 28 °C until removed from the thermal cycler.

5 µl of each 100 µl reaction was added to 5 µl loading dye (15% Ficoll, 0.05% Bromophenol blue, 0.05% Xylene cyanol in 1x AGB),* and electrophoresed on a 1.0% horizontal agarose gel (1x AGB running buffer) to check reaction specificity and yield. After electrophoresis, the gel was stained with 0.5 µg/ml ethidium bromide and photographed during illumination by short-wave ultraviolet light. Only those reactions containing 50 ng/µl or more (as gauged by comparison with 500 ng λ DNA cut with *Hind* III (GIBCO BRL Life Technologies Ltd.)) of the desired 3.067 kb product were retained (at 4 °C) for sequence analysis. Other reactions were repeated as required.

Preparation of Single-stranded DNA for Sequencing

Biotinylated PCR products of high yield and specificity were bound to streptavidin-coated magnetic beads (Dynal (UK) Ltd.) and denatured. Single-stranded DNA (ssDNA) for sequencing was then isolated magnetically.

Streptavidin-Biotin Binding

Four successful (i.e. high specific yield) 100 µl 3'-biotinylated PCR reactions per sample were combined and bound to streptavidin-coated magnetic beads ('Dynabeads') according to the manufacturer's suggested protocol, described below. The single (5'-biotinylated) PCR reaction was separately bound to another aliquot of beads.

A vol of Dynabeads (10 µg/ml) corresponding to one-half the total vol of the PCR reaction product to be bound (i.e. 0.5 vols) was washed before use to remove the 0.2%

---

* AGB, short for 'Agarose Gel Buffer', is better known as TAE Buffer (1x: 0.04 M Tris-Acetate, 0.001 M EDTA).

NaN_3 preservative. To wash, the beads were transferred to a 1.5 ml eppendorf tube and placed in one well of a six-well Dynal Magnetic Particle Concentrator (MPC-E) which held the Dynabeads immobilised while the non-magnetic supernatant was removed. The tube was then taken out of the MPC and the beads resuspended in 0.5 vols 1x Binding and Washing ('B & W') Buffer (5 mM Tris-Cl pH 7.5, 0.5 mM EDTA, 1 M NaCl). This wash was removed (using the MPC), and the beads finally resuspended in 1 vol 2x B & W Buffer.

Equal vols of washed Dynabeads and biotinylated PCR product were combined and allowed to incubate at room temperature, with occasional gentle mixing, for at least 15 min. After binding, the PCR-Dynabead complex was washed (as described above) in 1 vol 1x B & W Buffer. The bound product was then either stored at 4 °C, for up to 2 weeks, or made immediately into ssDNA.

Single-stranded DNA Preparation

The PCR-Dynabead complex was removed from its supernatant and resuspended in 0.2 vols of a freshly prepared 0.1 M NaOH solution. After alkaline denaturation for 10 min at room temperature, the beads (which are bound to only one strand of each PCR product because only one primer was biotinylated) were held by the MPC while the non-bound strand was removed with the supernatant. The remaining ssDNA was washed with 1 vol 0.1 M NaOH, 1 vol 1x B & W Buffer, and 1 vol TE Buffer pH 7.6, before final resuspension in 0.1 vol sterile H_20. This purified ssDNA, 5 µl of which is sufficient for one sequencing reaction, was stored at -20 °C.

DNA Sequence Analysis

Single-stranded DNA was sequenced using the dideoxy chain-termination method of DNA sequencing (Sanger, Nicklen, & Coulson, 1977). Reactions were electrophoresed, in parallel, on denaturing polyacrylamide gels.

Chain-Termination DNA Sequencing

Diploid DNA templates were sequenced by the chain-termination method using Sequenase Version 2 T7 DNA polymerase (United States Biochemical Corp.), according to the manufacturer's protocol.

Each sample was sequenced with 9 internal (i.e. lying within the region bounded by 5' AP and 3' AP) sequencing primers: eight 5' primers, designated 5' SP 1 to 5' SP 8, were used to sequence the whole of the 5'->3' strand (3' B'd template), while a single sequencing primer, 3' SP 8, was used to sequence the 5'-most portion of the 3'->5' strand (5' B'd template). The location of each of these primers within the 3 kb amplified region, as well as those of seven others (3' SP 1 to 3' SP 7), is shown in Figure A.1.

All sequencing reactions were performed in a similar fashion: 5 μl of ssDNA was combined with 1 to 3 μl (1 pmol/μl in most instances) sequencing primer (Table A.1), 2 ml 5x Sequenase Reaction Buffer (200 mM Tris-Cl pH 7.5, 100 mM $MgCl_2$, 250 mM NaCl), and H_20 to a final vol of 10 μl. The primer was annealed to the template by heating the mixture for 2 min at 65 °C and slowly cooling to less than 35 °C for approximately 30 min. The annealed primer-template mix was then kept on ice (for a maximum of 4 hours) until the labelling step.

To the ice-cold annealed DNA mixture was added 1 μl 0.1 M Dithiothreitol (DTT), 2 μl Labelling Mix (1.5 mM dGTP, 1.5 mM dCTP, 1.5 mM TTP), 0.5 μl [α-^{35}S] dATP (10 μCi/μl and 10 μM, 1000 Ci/mmol) (Amersham International, plc), and 2 μl Sequenase enzyme (13 u/μl in 20 mM KPO_4 pH 7.4, 1 mM DTT, 0.1 mM EDTA, 50% glycerol) diluted 1 to 8 with Enzyme Dilution Buffer (10 mM Tris-Cl pH 7.5, 5 mM DTT, 0.5 mg/ml BSA). This labelling reaction was incubated on ice for 2.5 to 5 min.

Following labelling, 3.5 μl aliquots of the reaction mixture were transferred to each of four microtitre plate ('Omniplates', Hybaid Ltd.) wells containing 2.5 μl of combined Termination/Extending mix (1.5 μl of ddNTP Termination Mix, containing 80 μM each of dGTP, dATP, TTP, dCTP, 50 mM NaCl, and 8 μM of the ddNTP [G, A, T, or C] and 1

Name	pmol Primer per 5 μl Template	Electrophoresis Time*	Bases Readable
5' SP 1	3	Short	60-320
5' SP 1a	2	Short	210-550
5' SP 2	2	Long	405-800
5' SP 3	15	Long	790-1180
5' SP 4	2	Long	1165-1560
5' SP 5	2	Long	1545-1940
5' SP 6	2	Long	1925-2340
5' SP 7	2	Long	2325-2700
5' SP 8	1	Long	2685-3050
3' SP 1	2	Short	2910-2550
3' SP 2	15	Long	2540-2230
3' SP 3	2	Long	2260-1860
3' SP 4	2	Long	1880-1480
3' SP 5	2	Long	1560-1160
3' SP 6	15	Long	1180-780
3' SP 7	2	Long	780-380
3' SP 8	2	Long	525-100

**Short = until bromophenol blue dye line has just run off the gel (approx. 2 hrs); Long = until the xylene cyanol dye line is 8 cm from the end of the gel (approx. 3.5 hrs).*

Table A.1 Reaction and electrophoresis specifications for the sequencing primers used in the study. The range of bases typically readable, given these conditions, is also shown.

ml of Sequence Extending Mix, containing 180 μM each of dGTP, dATP, TTP, dCTP, and 50 mM NaCl), pre-warmed to 42 °C in a Combi Thermal Reactor TR2 (Hybaid Ltd.). These termination reactions were allowed to incubate for 5 to 30 min at 42 °C before reactions were completed by the addition of 4 μl Stop Solution (95% formamide, 20 mM EDTA, 0.05% Bromophenol blue, 0.05% Xylene cyanol).

Reactions were most often performed in groups of 24 (one 96-well microtitre plate) and stored at -20 °C until gel electrophoresis.

Denaturing Gel Electrophoresis

Sequencing reactions were electrophoresed on 5% Long Ranger polyacrylamide gels, fixed, and dried to glass. Dried gels were then exposed to X-ray film.

'Long Ranger' (AT Biochem, from Hoefer Scientific Instruments) is the trade name for a chemically modified acrylamide monomer and co-monomer with novel cross-linker, which increases the number of readable bases per gel run by 30%. Electrophoresis with 5% Long Ranger (LR) gels followed the manufacturer's recommended protocol.

For a standard size sequencing gel (i.e. 20 x 40 cm), 40 ml of gel mix was prepared containing 16.8 g ultrapure Urea (GIBCO BRL, Life Technologies Inc.), 4.8 ml 10x TBE Buffer (0.9 M Tris-borate, 0.02 M EDTA), 4.ml LR gel solution (50% concentration), and H_20. Polymerisation was catalysed by the addition of 20 μl TEMED (N, N, N', N'-Tetramethylethylenediamine, Sigma Chemical Co., Ltd.) and 200 μl 10% Ammonium Persulfate (Aldrich Chemical Co., Ltd.). The polymerising mixture was poured between two glass plates, one siliconised (using SurfaSil™ siliconising fluid from Pierce & Warriner) and one abraded and treated with a 15: 150: 5000 μl solution of 'Bind-silane' (γ-methacryloxypropyltrimethoxysilane, Sigma Chemical Co., Ltd.): 10% Acetic Acid: Ethanol (for glass-acrylamide binding), spaced at a thickness of 0.4 mm, and set with a 48-well sharkstooth comb (2.9 mm point-to-point, BRL, Life Technologies Inc.). LR gels polymerised in 30 min to 1 hour and were typically prepared 2 to 24 hours prior to use.

Each polymerised gel was placed in a vertical slab gel electrophoresis tank (Scotlab), and pre-run for 10 to 20 min in 0.6 x TBE Buffer at 35 W, maximum current and voltage. The loading well was rinsed free of excess urea, and the comb inserted, just prior to sample loading. Samples (12 sets of reactions per gel) were heated to 80 °C for 5 min and transferred immediately to ice to prevent strand renaturation. 1.5 µl of each termination mix was loaded to each well. Reactions were loaded in parallel, with ddGTP termination reactions from samples 1 to 12 in adjacent lanes, followed (left to right) by ddATP 1 to 12, dTTP 1 to 12, and ddCTP 1 to 12 (see Figure 2.3). Samples underwent electrophoresis at 35 W for 2 to 3.5 hours, depending upon the primer used (Table A.1).

When electrophoresis was complete, the top (siliconised) plate was removed and the back plate, with gel attached, was fixed by soaking for 2x 30 min in 500 ml 20% ethanol, 10% glacial acetic acid. The fixed gel was dried to the plate at 100 °C for at least 15 min.

Once cool, the dried gel-plate was exposed directly to either 'fast' XAR-5 (Kodak, from Sigma Chemical Co., Ltd.) or 'slow' RX (Fuji) X-ray film (35 x 43 cm) for 1 to 7 days in a light-tight autoradiograph cassette. Films were developed automatically, using the X-ograph Varispeed X150 processor at the Institute of Molecular Medicine (IMM), Oxford. Following autoradiography, the plate was soaked in H_20 to re-hydrate the LR gel for removal and disposal.

Allele-specific Amplification for Linkage Determination

Allele-specific primers were used to amplify DNA from one chromosome of each heterozygote. Allele-specific PCR products were then sequenced to directly determine the relevant sequence haplotype. Most of the corresponding haplotype was inferred by comparison.

General Strategy

Once the nature and position of heterozygous nucleotide positions were identified in each sample, Amplification Refractory Mutation System, or ARMS, primers (Newton et al. 1989) were chosen for allele-specific amplification of regions containing unresolved linkage relationships.

For the majority of samples, only a small number of ARMS primers were required. These were designed to discriminate sites of high heterozygosity, often with proximity to either the 5' or 3' end of the 3.0 kb region (Table A.2). In each case, the ARMS primer was chosen such that its 3'-most nucleotide fell at the heterozygous site and was homologous to one of the two bases present. In addition, the primer was synthesised with a destabilising internal (position -3) mismatch (Kwok et al. 1990), so that *Taq* polymerase extension would only occur if there was perfect homology at the 3'-most base. Under optimal annealing conditions, the resulting PCR product was chromosome-specific because only DNA lying between the homologous base and the reverse primer was amplified.

The actual ARMS primer used was decided on a sample-specific basis, as different sites were heterozygous in different individuals. Where possible, the furthest flanking heterozygous site (of a series of segregating sites) was targeted, so that only a single allele-specific amplification product was required. The appropriate complementary biotinylated amplification primer (5' B'd AP or 3' B'd AP) was used in all cases, to biotinylate the product for sequencing.

Where the diploid sequence was unambiguous, one chromosome was specifically amplified (the sequence haplotype of the second allele was inferred from comparison of the directly-determined haploid and diploid sequences). At the 5' end of the 3.0 kb region, however, length polymorphism made interpretation of the diploid sequence difficult, hence both chromosomes were separately amplified for sequencing, using reciprocal ARMS primers (e.g. ARMS 5'A and antiARMS 5'A).

Name	Site Amplified	Base Amplified (as on 5'->3' strand)	Conditions Ann. Temp. (°C)	Cycles
ARMS 5'A	145	C	65	40
AntiARMS 5'A	145	T	65	40
ARMS 5'ßs	917	T	68	30
AntiARMS 5'ßs	917	A	68	30
ARMS 3'A	2636	C	60	40
AntiARMS 3'A	2636	A	65	40
ARMS 3'B	2945	G	68	45
ARMS 3'C	145	C	65	40
ARMS 3'D	2792	T	60	40
ARMS 3'ßs	917	T	68	30
AntiARMS 3'ßs	917	A	68	30

Table A.2 ARMS primers used for chromosome-specific PCR amplification, showing the site and base amplified, as well as the reaction conditions required.

Allele-Specific PCR Amplification and Sequencing

Genomic DNAs were PCR-amplified, as previously described, using 20 pmol of the appropriate ARMS primer per 100 μl reaction. PCR conditions* for each ARMS primer, optimised for amplification specificity, are given in Table A.2. Reactions containing 100 ng/μl or more were retained for sequence analysis. The number of 100 μl ARMS reactions required per sample depended on the number of sites to be sequenced (50 μl of PCR product was needed per sequencing reaction). Only previously heterozygous positions needed to be re-sequenced in the ARMS product.

Single-stranded DNA preparation and DNA sequence analysis was as described for diploid PCR amplifications. However, when samples had been amplified with ARMS 5'A or antiARMS 5'A, 5' SP 1 was replaced by 5' SP 1a (Figure A.1) which, unlike 5' SP 1, annealed within the amplified region. Similarly, 5' biotinylated PCR products were sequenced, as appropriate, with 3' sequencing primers.

Restriction Enzyme Analysis of Flanking Sites

Flanking restriction fragment length polymorphisms (RFLPs) were investigated by PCR amplification and restriction enzyme digestion of the relevant polymorphic sites.

PCR Amplification of Polymorphic Sites

Nine polymorphic restriction enzyme sites were PCR-amplified from genomic DNA using seven pairs of amplification primers.

Seven separate genomic regions, spanning a region over 50 kb in length, were amplified by PCR, using primers designed by J. Old (personal communication), Sutton,

---

* These reaction conditions are abbreviated. The cycles described were each preceded by an initial cycle of 4 min at 94 °C, 2 min at the given annealing temperature, and 5 min at 72 °C. A final cycle, with longer extension time of 10 min at 72 °C, followed.

Bouhassira, & Nagel (1989) and Fullerton & Clegg (1994) (Table A.3). The approximate positions of the 14 amplification primers used, as well as the 9 polymorphic restriction enzyme sites they encompass, are given in Figure A.2.

PCR amplification reactions were as described for the 3.067 kb ß-globin fragment, except that 20 pmol primer/100 µl was normally used and reaction vols were typically 50 µl (in the case of the Kpn I product) or 25 µl. In addition, for products less than a kilobase in length (all except Kpn I, see Table A.3) gp32 was not added to the PCR. Reaction conditions* for each set of amplification primers are given in Table A.3.

5 µl of each reaction was checked by horizontal gel electrophoresis, as previously described. Any suboptimal reactions (i.e. no product, or non-specific product) were repeated. Others were retained at 4 °C for restriction enzyme digestion.

Restriction Enzyme Digestion

PCR amplification products were digested with the appropriate restriction endonuclease to identify the presence or absence of the enzyme recognition site in samples.

10 µl of PCR product (500-1000 ng) was digested with 5 to 10 units of enzyme (Table A.3) in a final vol of 20 µl, using the restriction enzyme buffer provided by the manufacturer (Boehringer Mannheim Biochemica, New England Biolabs). Because the Kpn I PCR product contained three polymorphic restriction enzyme sites, 10 µl aliquots of that product were separately digested with either *Hpa* I, *Hind* III, or *Bam* HI. Digests were incubated at 37 °C for 4 to 14 hours.

After digestion, 10 µl of AGB loading dye was added and the samples electrophoresed on 1.0% (Kpn I digests) or 2.5% horizontal agarose gels. The gels were stained and photographed under UV as previously described. Fragment sizes resulting from restriction enzyme digestion are given in Table A.3.

---

* Reaction conditions are abbreviated, as previously described.

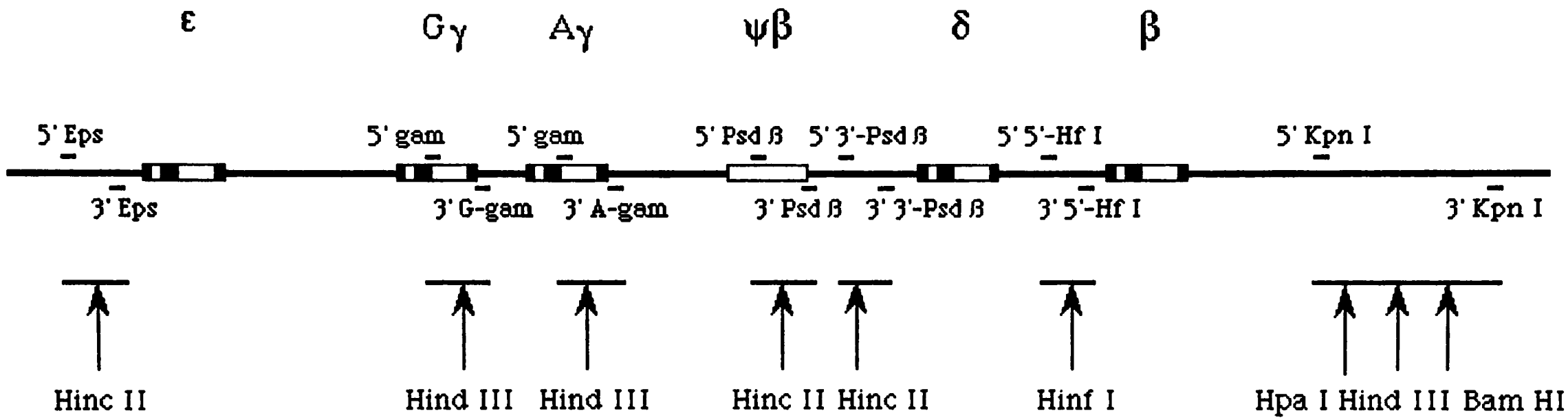


Figure A.2 The human ß-globin gene cluster, showing the annealing location of the RFLP haplotyping primers. The nature and approximate position of the restriction enzyme sites within each of the seven fragments amplified is shown below.

Name	Primer/100 µl (pmol)	Conditions Ann. Temp. (°C)	Cycles	Product Size	Enzyme	Cut Fragment Sizes	Source
5' Epsilon	40	60	30	760	Hinc II	446, 314	J. Old
3' Epsilon	40						
5' G/A-gamma	20	60	30	770	Hind III	430, 340	Sutton et al.
3' G-gamma	20						
5' G/A-gamma	20	65	30	760	Hind III	400, 360	Sutton et al.
3' A-gamma	20						
5' Pseudo ß	20	60	30	700	Hinc II	360, 340	Sutton et al.
3' Pseudo ß	20						
5' 3'-Psd. ß	20	60	30	613	Hinc II	593, 120	Sutton et al.
3' 3'-Psd. ß	20						
5' 5'-Hinf I	20	60	30	438	Hinf I	238, 200	Sutton et al. (5') & Fullerton (3')
3' 5'-Hinf I	20						
5' Kpn I	20	65	35	4259	Hpa I	3700, 559	
3' Kpn I	20			4259	Hind III	2517, 1742	Fullerton & Clegg
				4259	Bam HI	2920, 1339	

Table A.3 PCR primers and reaction conditions for ß-globin gene cluster RFLP haplotyping.

Samples Investigated

The survey of DNA sequence variation described here would not have been possible without the generosity of a number of researchers, who made available a significant body of samples for study. The source of each group of DNA samples investigated is acknowledged below. This is followed by a catalogue of the specific samples used in the study.

Sources of Samples

Samples investigated here included 24 normal (i.e. no known ß-haemoglobinopathy) individuals born on the island of Espiritu Santo in northern Vanuatu, island Melanesia (courtesy of M. Ganczakowski, IMM, Oxford), 12 ß-thalassaemia (IVS 1: 5, G->C) heterozygotes from the island of Maewo, northern Vanuatu (also from M. Ganczakowski), 12 normal Papua New Guinea (PNG) individuals from the Eastern Highlands Province (courtesy of D. R. Higgs, IMM, Oxford), and 2 Hb S homozygotes, 22 Hb S heterozygotes, and 3 normal individuals from the Gambia, West Africa (courtesy of A. V. S. Hill, IMM, Oxford).

In addition, data obtained from 12 normal individuals and 2 ß-thalassaemia (IVS 1:5, G->C) heterozygotes from Palembang, Sumatra in Indonesia (courtesy of A. S. M. Sofro, Gadjah Mada University, Indonesia; sequenced by J. Bond, IMM, Oxford) were also considered in the thesis' analyses.

Samples Analysed

Table A.4 gives the supplying investigator's identification number, and the subsequently adopted label, for the 89 samples used in the study. Samples are grouped under each location according to their haematological status.

Adopted Name (current study)	Sample Name (supplying investigator)	Adopted Name	Sample Name	Adopted Name	Sample Name	Adopted Name	Sample Name
NORTHERN VANUATU (POP 1)		**MAEWO (POP 2)**		**INDONESIA (POP 4)**		**GAMBIA (POP 5)**	
Normals		**ß-thalassaemia Heterozygotes**		**Normals**		**Normals**	
P1-1	SB 130	P2-1	MB 26	P4-1	PLB 013	P3-3	GT 1666
P1-2	SB 148	P2-2	MB 40	P4-2	PLB 015	P3-13	GM 260
P1-3	SB 152	P2-3	MB 57	P4-3	PLB 020	P3-S1	GM 255
P1-4	SB 101	P2-4	MB 60	P4-4	PLB 024		
P1-5	SB 124	P2-5	MF 66	P4-5	PLB 081	**Hb S Homozygotes**	
P1-6	SB 125	P2-6	MB 84	P4-6	PLB 085		
P1-7	SB 127	P2-7	MB 169	P4-7	PLB 092	P3-S3	GM 385
P1-8	SB 145	P2-8	MF 73	P4-8	PLB 093	P3-S4	GM 94
P1-9	SB 159	P2-9	MF 88	P4-9	PLB 103		
P1-10	SB 81	P2-10	MF 155	P4-10	PLB 114	**Hb S Heterozygotes**	
P1-11	SB 93	P2-11	MF 172	P4-11	PLB 047		
P1-12	SB 97	P2-12	MM 39	P4-12	PLB 070	P3-1	GT 1649
P1-13	SB 106					P3-2	GT 1661
P1-14	SB 107			**ß-thalassaemia Heterozygotes**		P3-4	GM 15
P1-15	SB 109	**PAPUA NEW GUINEA, EASTERN HIGHLANDS (POP 3)**				P3-5	GM 57
P1-16	SB 110			P4-13	RSS-048	P3-6	GM 103
P1-17	SB 111	**Normals**		P4-14	RSS-007	P3-7	GM 120
P1-18	SB 118					P3-8	GM 155
P1-19	SB 122	P2-13	4563			P3-9	GM 205
P1-20	SB 123	P2-14	4567			P3-10	GM 207
P1-21	SB 155	P2-15	4570			P3-11	GM 215
P1-22	SB 158	P2-16	4573			P3-12	GM 233
P1-23	SB 168	P2-17	4577			P3-14	GM 326
P1-24	SB 175	P2-18	4579			P3-15	GM 330
		P2-19	4583			P3-16	GM 355
		P2-20	4587			P3-17	GM 499
		P2-21	4589			P3-18	GT 1374
		P2-22	4590			P3-19	GT 1450
		P2-23	4593			P3-20	GT 1488
		P2-24	4597			P3-21	GT 1521
						P3-22	GT 1547
						P3-23	GT 1583
						P3-24	GT 1587

Table A.4 Summary of the samples used in the analysis of ß-globin allelic sequence diversity.

Oligonucleotides Used in the Study

Oligonucleotides were used extensively throughout this study. As described above, oligonucleotide primers were required for the amplification of genomic template for sequencing, as well as for DNA sequencing itself. Specially designed primers permitted allele-specific amplification for linkage determination, and a further set of oligonucleotides were used to PCR-amplify adjacent polymorphic restriction enzyme sites for haplotyping.

A brief outline of the method by which the oligonucleotide primers used in this study were chosen, and subsequently prepared for use, is followed by an extended table (Table A.5) giving the name, date of deprotection, sequence, stock concentration, and working concentration of each primer.

Oligonucleotide Primer Design

Except where otherwise stated, the oligonucleotides used in this study were chosen with the help of the computer program PRIMER, Version 0.5 (Lincoln, Daly, and Lander, 1991). This program predicted primer annealing temperature and specificity, as well as providing estimates of self-homology and homology to complement oligonucleotides.

Oligonucleotide Synthesis and Deprotection

All oligonucleotide primers used in this study were prepared at the MRC Molecular Haematology Unit, IMM, Oxford, on a 381A DNA synthesiser (Applied Biosystems Inc.), using standard phosphoramidite chemistry.

Biotinylation of the 5' and 3' 3.0 kb amplification primers (labelled 'B'd' in Table A.5)

Name	Date	Sequence 5'->3'	Stock [ ] (pmol/µl)	Working [ ] (pmol/µl)
Template Amplification Primers				
5' AP	6.90	AGT CCA GGC AGA AAC AGT TAG ATG TC	355.8	10
3' AP	8.5.92	GCC CAG TCC TTC CAA AGC AGA C	239.5	10
Biotinylated Amplification Primers				
5' B'd AP	15.5.92	B- AGT CCA GGC AGA AAC AGT TAG ATG TC	444.3	10
3' B'd AP	15.5.92	B- GCC CAG TCC TTC CAA AGC AGA C	348.5	10
Sequencing Primers				
5' SP 1	15.5.92	GAT GTC CCC AGT TAA CCT CC	462.5	1
5' SP 1a	20.8.93	GTA ACT AAT GCA CAG AGC AC	810.2	1
5' SP 2	19.8.92	GAT ATG CTT AGA ACT GAG GTA GAG	626.0	1
5' SP 3	15.5.92	TGA AGT CCA ACT CCT AAG CC	504.5	5
5' SP 4	25.4.92	GAC TCT TGG GTT TCT GAT AG	885.8	1
5' SP 5	15.5.92	TAG AAT GGG AAA CAG ACG	445.4	1
5' SP 6	15.5.92	CAG GGT AAT TTT GCA TTT G	492.7	1
5' SP 7	15.5.92	GTC TGT GTG CTG GCC CAT C	498.6	1
5' SP 8	17.2.92	CCA TGA AAG AAG GTG AGG C	545.8	1
3' SP 1	25.11.92	GCC CCC TGT TTT TCC TTC	775.5	1
3' SP 2	6.90	GCA TAG GCA TCA GGG CTG TTG	333.7	5
3' SP 3	30.10.92	GTT GGA CTT AGG GAA CAA AGG	395.1	1
3' SP 4	30.10.92	GCT ATT GCC TTA ACC CAG	479.6	1
3' SP 5	29.9.92	GTA ATG TAC TAG GCA GAC TG	588.6	1
3' SP 6	30.10.92	CTC AGT GTG GCA AAG GTG	566.8	5
3' SP 7	30.10.92	CAT GGT GTC TGT TTG AGG	610.4	1
3' SP 8	30.10.92	CTC TTG GCC CCA TAC CAT C	581.3	1
ARMS Amplification Primers				
ARMS 5'A	30.10.92	GAC TGC ATT AAG AGG TCT CTA GTT TTG TAC	540.0	20
AntiARMS 5'A	13.11.92	GAC TGC ATT AAG AGG TCT CTA GTT TTG TAT	304.5	20
ARMS 5'ßs	16.11.93	CAT GGT GCA CCT GAC TCA TGT	622.0	20
AntiARMS 5'ßs	4.10.93	CAT GGT GCA CCT GAC TCA TGA	696.8	20
ARMS 3'A	16.9.92	CCT TTT CTG AGG GAT GAA TAA GGC CTA G	392.4	20
AntiARMS 3'A	30.10.92	CCT TTT CTG AGG GAT GAA TAA GGC CTA T	266.3	20
ARMS 3'B	13.11.92	CTC ACA GGA CAG TCA AAC CCT GC	467.9	20
ARMS 3'C	25.11.92	CTT ATT AGG TTT TGG GAA ACA ATA GG	722.3	20
ARMS 3'D	10.12.92	CTC TAA GCA AGA GAA CTG AGT GCA GA	741.7	20
ARMS 3'ßs	4.10.93	GCA GTA ACG GCA GAC TTC ACC T	795.0	20
AntiARMS 3'ßs	4.10.93	GCA GTA ACG GCA GAC TTC ACC A	937.7	20
Haplotyping Amplification Primers				
5' Epsilon	23.9.92	TCT CTG TTT GAT GAC AAA TTC	619.3	20
3' Epsilon	23.9.92	AGT CAT TGG TCA AGG CTG ACC	742.1	20
5' G/A-gamma	23.9.92	TGC TGC TAA TGC TTC ATT ACA A	550.4	20
3' G-gamma	23.9.92	AAG TGT GGA GTG TGC ACA TGA	664.7	20
3' A-gamma	16.9.92	TAA ATG AGG AGC ATG CAC ACA C	517.3	20
5' Pseudo ß	16.9.92	GAA CAG AAG TTG AGA TAG AGA	536.5	20
3' Pseudo ß	16.9.92	ACT CAG TGG TCT TGT GGG CT	622.3	20
5' 3'-Psd. ß	16.9.92	TCT GCA TTT GAC TCT GTT AGC	501.9	20
3' 3'-Psd. ß	16.9.92	GGA CCC TAA CTG ATA TAA CTA	643.4	20
5' 5'-Hinf I	19.8.92	CTA CGC TGA CCT CAT AAA TG	675.5	20
3' 5'-Hinf I	17.2.92	TGG GGT AAT CAG TGG TGT C	793.7	20
5' Kpn I	7.10.92	CAT GCC AAA TTG TAA ACG ACC ATC AAG G	500.5	20
3' Kpn I	19.8.92	GAT AAT GGC TCT ACG GAT GTG TGA GAT C	820.9	20

Table A.5 Principal oligonucleotide primers used in the study.

was achieved by re-synthesis with direct incorporation of a biotin phosphoramidite (DMT-Biotin-C6-PA, Cambridge Research Biochemicals, Ltd.) at the 5' end of each oligonucleotide. Oligonucleotides biotinylated during synthesis are treated like any other primers (see below) and bind streptavidin more readily than primers labelled with biotin after synthesis (S. L. Thein, personal communication).

Following synthesis, each oligonucleotide was 'deprotected' prior to use. Primer (bound to beads) was removed from the synthesis column and incubated at 55 °C for 6 or more hours in 0.5 ml NH_4OH. Deprotected primer was then precipitated with 0.1 vol 3 M NaOAc pH 5.8 and 1 ml 100% ethanol, spun at 12,000 rpm for 10 minutes, washed in 85% ethanol, spun, and finally resuspended in 200 µl sterile H_20. Oligonucleotide concentration was determined by taking the optical density (OD) of a 1 to 500 dilution at 260 nm wavelength ($\{[OD_{260} \times 0.037] / [330 \times \text{Oligo Length}]\} \times 10^6 = \text{pmol}/\mu\text{l}$). All oligonucleotides were stored at -20 °C.

Oligonucleotide Primers Used

Table A.5 catalogues the oligonucleotide primers used in the principal analyses described above. Other details, including location of annealing as well as optimal annealing temperature and/or cycling conditions, are described earlier in the chapter. Less commonly used primers are given in Appendix B.

Appendix B

Supplementary Methods and Materials

Materials and methods described here refer to protocols which were explored, but not subsequently adopted for use in the larger study (Chapter 2).

Simultaneous Cloning Protocol

2.6 kilobase (kb) PCR products were amplified from 64 samples using specific combinations of tagged amplification primers (APs). Aliquots of each PCR product were pooled, digested with restriction enzymes, and purified by gel electrophoresis onto DEAE membrane. This insert DNA was ligated to M13 mp18 or mp19 and *E. coli* transfected with the recombinants. Single-stranded DNA was prepared from positive (colourless) colonies. Either end of the insert in each single-stranded DNA (ssDNA) preparation was sequenced to determine clone origin. A subset of mp18-cloned inserts were sequenced by the dideoxy chain-termination method.

Tagged Amplification Primers

Sixteen oligonucleotide PCR primers, homologous to the same sequences as 5' AP (Table A.5) and 3' (2.6 kb) AP (5'-GCATAGGCATCAGGGCTGTTG-3', anneals 2641 basepairs [bp] 3' of 5' AP), but containing at their 5' ends artificial sequences for restriction enzyme recognition and insert identification, were synthesised. The sequence of the four 5' APs was 5'-ACATCTGCAG**X**AGTCCAGGCAGAAACAGTTAGAT GTC-3', where **X** = G, A, T, or C (*Pst* I site underlined), and the sequence of the sixteen 3' APs was 5'-ACGTGAGCTC**XX**GCATAGGCATCAGGGCTGTTG-3', where **XX** = GG, AA, TT, CC, GA, GT, GC, AG, AT, AC, TG, TA, TC, CG, CA, or CT (*Sac* I site

underlined).

PCR Amplification

PCR amplification was as described for the 3.0 kb PCR (Appendix A) except that 20 pmol of each primer was used and reaction conditions were 1 cycle of 3 min at 94 °C, 2 min at 60 °C, 5 min at 72 °C, 30 cycles of 1 min at 94 °C, 2 min at 65 °C, 5 min at 72 °C, and a final cycle of 1 min at 94 °C, 2 min at 65 °C, and 10 min at 72 °C. A unique combination of 5' and 3' APs was used for each sample amplified. Product concentration and specificity was checked by agarose gel electrophoresis. Reactions containing at least 20 ng/μl of the 2.6 kb product were retained for cloning. Amplifications were repeated (with different samples, as necessary) until 64 products of the desired concentration were obtained.

Insert Preparation

5 μl of each of 64 uniquely amplified PCR products were combined. The pooled DNA was precipitated with 0.1 volume (vol) 4 M LiCl, 3 vols 95% ethanol and reconstituted in 16 μl TE pH 7.6. Insert DNA was digested with 10 units *Pst* I and 10 units *Sac* I in the buffer supplied by the manufacturer (Intl. Biotechnologies, Inc.) at 37 °C for 1 hr.

Digested DNA was electrophoresed onto DEAE-cellulose membrane (Schleicher & Schuell), as described by Sambrook, Fritsch, & Maniatis (1989). Following electrophoresis, the membrane was incubated at 65 °C for 30 min in high salt buffer (1 M NaCl, 50 mM Tris-Cl pH 8, 10 mM EDTA pH 8) to elute the DNA. The eluate was extracted once with phenol:chloroform:isoamyl alcohol (25:24:1 v/v, pH 7) and once with chloroform:isoamyl alcohol (24:1), then precipitated with 0.2 volumes 10 M ammonium acetate and 2 volumes 100% ethanol. DNA was resuspended in 10 μl TE pH 8.

DNA Cloning

Purified pooled insert was cloned into M13 vectors according to the manufacturer's protocol (Boehringer Mannheim). 100 ng insert DNA was ligated to 10 ng M13 mp18 or mp19 (previously digested with *Pst* I and *Sac* I). 300 µl competent *E. coli* (strain DH5αF') cells, prepared by incubation with ice-cold $CaCl_2$, were transfected with 5 µl ligation mix, and plated in IPTG/X-gal-containing top agar. Plates were incubated at 37 °C overnight.

ssDNA Preparation

96 colourless plaques per vector were isolated and used to inoculate fresh DH5αF' cultures. After incubation at 37 °C for 6 hours, *E. coli* cells were removed by centrifugation and phage isolated from the supernatant by precipitation with 2.5 M NaCl, 20% (w/v) polyethylene glycol 6000 (BDH Chemicals Ltd.). The phage pellet was phenol extracted, precipitated with 4 M LiCl and 95% ethanol, and resuspended in 50 µl TE pH 8. DNA size and concentration was checked by agarose gel electrophoresis.

Clone Identification

The 5' and 3' ends of each cloned insert were sequenced to identify the primers' base tags, and thus, the sample of origin. The universal M13 sequencing primer -40 to the polylinker (United States Biochemical Corp.) was used to sequence the 5' end of inserts cloned in mp18 and the 3' end of mp19 inserts. Additional, internal, sequencing primers were synthesised to sequence the 3' end of inserts (5'-CCATGAAAGAAGGTGAGGC-3') or insert 5' ends (5'-TGGGGTAATCAGTGGTGTC-3'). Sequencing was as described in Appendix A, except that 3.5 µl of the ssDNA template was used and 1 µl Mn Buffer (US Biochemical) was added to each labelling reaction to highlight sequences close to the primer. Reactions were electrophoresed on 6% denaturing polyacrylamide gels (20:1 acrylamide/N, N'-methylene bisacrylamide, 7 M urea) in 1 x TBE buffer (Appendix A) and visualised by autoradiography.

DNA Sequence Analysis

The first 600 bp of 36 mp18-cloned samples were sequenced by the dideoxy chain termination method. Sequencing was as described for the clone identifications, except that no Mn Buffer was added. Templates were sequenced with the -40 universal primer, and with 3 other internal sequencing primers: mp18-5'1 (5'-GACATAATTTATTAGCATG C-3'), mp18-5'2 (5'-CTCTTCCACTTTTAGTGC-3'), and mp18-5'3 (5'-GGCCAATCTACTC CCAGG-3').

SSCP Analysis of Four-Cutter Digested PCR Product

2.6 kb PCR products, isotopically-labelled during amplification, were digested separately with one of five four-cutter restriction enzymes. Aliquots of these digests were then either denatured or not denatured, and electrophoresed through non-denaturing polyacrylamide gels. Non-denatured fragments were scored for restriction fragment length polymorphisms (RFLPs) resulting either from enzyme recognition site modification or length differences. Denatured fragments were scored for the presence or absence of SSCP mobility shifts.

PCR Amplification

PCR amplification was as described for the 3.0 kb PCR (Appendix A), except that a different 3' AP, the 3' (2.6 kb) AP described above, was used and 1 µl of [α-^{32}P] dCTP (10 mCi/ml and 3000 Ci/mmol) was added for *in vitro* isotopic labelling (Orita et al., 1989b). Following amplification, the labelled 2.6 kb fragment was separated from unincorporated nucleotides and primers by 'gene-cleaning', i.e. reversible adsorption of PCR product, but not contaminants or primers, to a specially-formulated silica matrix (Bio 101, Inc., La Jolla, California), and resuspended in 50 µl H_20.

Four-cutter Digests

8 μl aliquots of purified PCR product were digested with 10 units of the appropriate restriction enzyme (*Sau* 3A I, *Dde* I, *Hinf* I, *Alu* I, or *Hae* III) in the buffer provided by the manufacturer (Boehringer Mannheim). After overnight incubation at 37 °C, digest volumes were increased to 25 μl, and each was extracted with phenol:chloroform:water (70:20:10 v/v) and chloroform.

Sample Preparation

5 μl of each extracted digest was combined with 1 μl 6x PAGE loading dye (0.25% bromophenol blue, 0.25% xylene cyanol, 15% Ficoll) for non-denaturing (ND) gel electrophoresis of the restriction enzyme fragments. Another 2 μl aliquot of each extracted digest was separately combined with 2 μl 0.1% SDS, 10 mM EDTA and 1 ml of 95% formamide, 20 mM EDTA, 0.05% (w/v) bromophenol blue, 0.05% xylene cyanol for SSCP electrophoresis. The latter aliquots were denatured by heating to 90 °C for 3 min, then quick cooled on ice, just prior to loading.

Gel Electrophoresis

2 μl of each sample (ND or SSCP) was electrophoresed on 6% (*Sau* 3A, *Hae* III) or 8% (*Dde* I, *Hinf* I, *Alu* I) polyacrylamide gels (49:1 acrylamide/N, N'-methylene bisacrylamide, 10% glycerol) pre-run for 1 hr in 1x TBE buffer (Appendix A). Samples were electrophoresed at 200 V, 180 mA, 15 W for 15 to 20 hours. Each gel was fixed by soaking in 7.5% acetic acid for 15 min, dried to its glass plate, and exposed to fast-speed x-ray film for 2 to 7 days.

Taq Polymerase Extension of Carbodiimide-Modified PCR Product

2.6 kb PCR products were denatured and reannealed to generate DNA heteroduplexes between the alleles present in each sample. Heteroduplexed DNAs were reacted with

carbodiimide (CDI), extracted, and heated to remove residual chemical. The CDI-modified DNAs were then used as template in a one-cycle PCR reaction, using only a single oligonucleotide primer (either one of the original amplification primers or an internal primer) for *Taq* polymerase extension. Reaction products were electrophoresed on denaturing polyacrylamide gels to determine the length of the amplified ssDNA.

PCR Amplification

PCR amplification was as described for the RFLP-SSCP protocol, except that [α-^{32}P] dCTP was not added. Product concentration and specificity was checked by agarose gel electrophoresis (Appendix A). Each successful 2.6 kb PCR product was extracted with phenol:chloroform:water (70:20:10 v/v) and chloroform and precipitated with 0.5 volumes 7.5 M ammonium acetate and 2 volumes ethanol. Precipitated DNA was resuspended in 50 μl TE, pH 7.4.

Heteroduplex Formation

20 μl reconstituted PCR product was combined with 10 μl hybridisation buffer (3 M NaCl, 35 mM $MgCl_2$, 30 mM Tris-Cl, pH 7.4) and 70 μl sterile H_20. Samples were denatured by heating to 100 °C for 10 min, and then incubated at 42 °C overnight for strand reannealing. Following incubation, heteroduplexed DNA was extracted once with chloroform, then ethanol precipitated. Heteroduplexed DNA was resuspended in 60 μl TE, pH 7.4.

CDI Modification

26 μl of each heteroduplexed sample was combined with 4 μl 1 M sodium borate, pH 8 and 10 ml 200 mM CDI (1-cyclohexyl-3-[2-(4-morpholinyl)ethyl]-carbodiimidemetho-*p*-toluenesulfonate, Fluka Chemical Corp.), and incubated at 30 °C for 3 hrs in siliconised tubes (CDI-treated DNA adheres to non-siliconised glass and plastic). Following incubation, 40 μl 10 mM sodium phosphate, pH 7 and 40 μl 7.5 M ammonium acetate was added. The samples were then extracted three times with isoamyl alcohol and

ethanol precipitated. CDI-modified DNA was resuspended in 26 µl TE, pH 7.4, and heated to 100 °C for 5 min to destroy residual CDI.

One-cycle PCR

4 µl of the CDI-modified template was placed with 5 pmol of one of the following PCR primers: AP 1 or 5' AP (Table A.5), AP 2 or 3' AP (described above), AP 3 or 3' CHA 3A (5'-TTGATGCACTAAAAGTGGAAGAG-3', anneals 500 bp 3' of 5' AP), and AP 4 or 5' ß7 (5'-CAATGTATCATGCCTCTTTGCAC-3', anneals 700 bp 5' of 3' AP). Each template/primer mixture was amplified in a total volume of 10 µl (50 mM KCl, 10 mM Tris-Cl pH 8.4, 1.5 mM $MgCl_2$, 0.001% gelatin, 1.25 mM each dNTP) containing 0.25 ml [α-^{32}P] dCTP and 0.2 units *Taq* DNA polymerase. Reactions underwent 1 cycle of 4 min at 94 °C, 2 min at 60 °C, and 5 min at 72 °C before being stopped by the addition of 6 ml loading dye (95% formamide, 20 mM EDTA, 0.1% [w/v] bromophenol blue and 0.1% xylene cyanol). Similar amplification of non-CDI-treated DNAs was also performed.

Gel Electrophoresis

8 µl of each sample was heated to 94 °C for 2 min, quick cooled on ice, and loaded to a 6% denaturing polyacrylamide gel (20:1 acrylamide/N, N'-methylene bisacrylamide, 7 M urea) pre-run in 1 x TBE buffer (Appendix A). Each gel was run at 30 W, full current and voltage, for 1 to 1.5 hrs. Following electrophoresis, the gel was fixed in 10% acetic acid for 0.5 hrs, dried to its glass plate, and exposed to fast-speed x-ray film for 1 to 7 days.

Appendix C

Phylogenetic Analysis of ß-Globin Sequence Haplotypes

Figure C.1 shows one possible tree of genealogical relationship for the 53 ß-globin sequence haplotypes, constructed using the tree-building computer program PAUP, Phylogenetic Analysis Using Parsimony (Swofford, 1991). The tree is a consensus of 5 different trees identified during a general heuristic search using simple stepwise addition and tree bisection-reconnection branch swapping. All insertion/deletion polymorphisms, non-silent variants, and nucleotide polymorphisms found in the 5' flanking DNA were excluded from consideration in the tree-building exercise.

The tree presented in Figure C.1 should be regarded as representative only of the larger genetic relationships among the sequence haplotypes and not a definitive description of the pattern of allelic divergence at the ß-globin locus. Reliable inference of ß-globin sequence haplotype genealogy awaits more detailed consideration of these, and related, data.

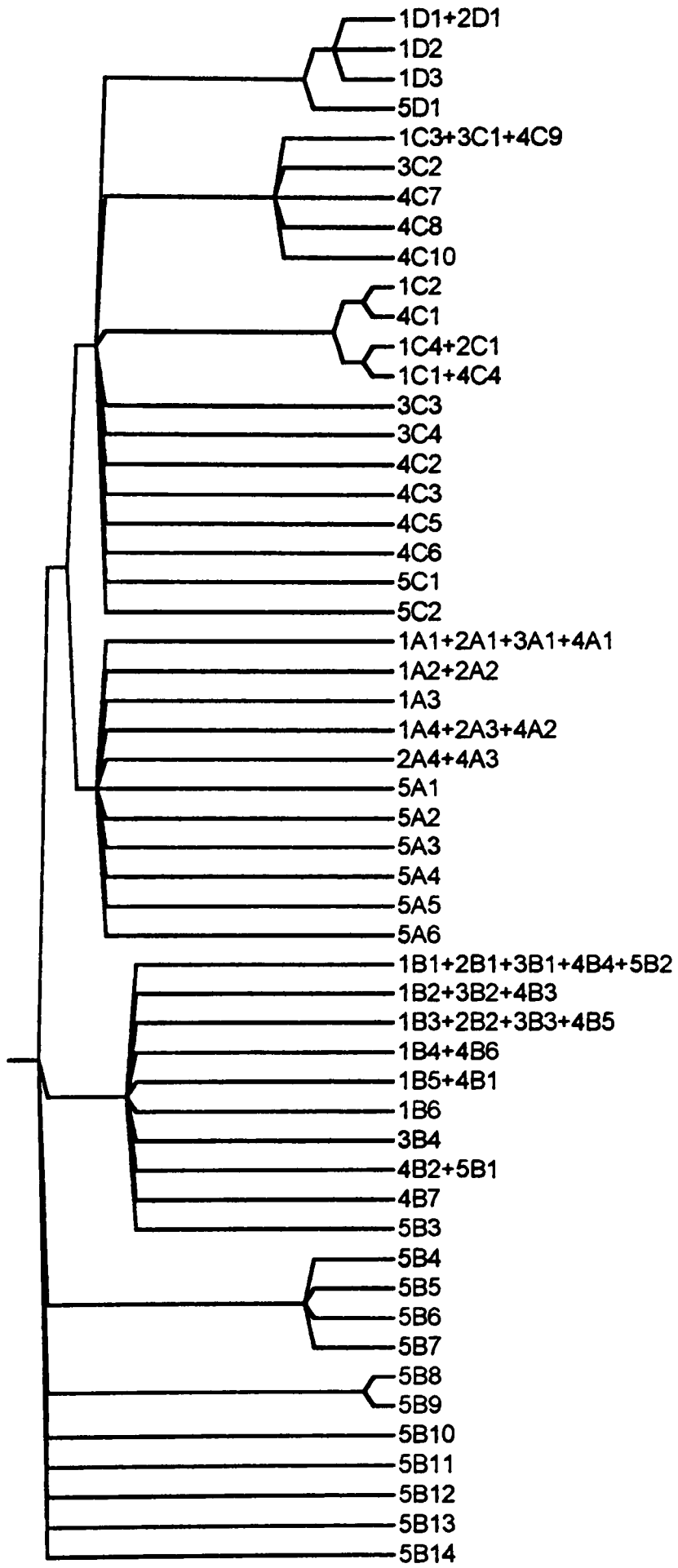


Figure C.1 A genealogical tree showing the hypothetical ancestral relationship of the 53 ß-globin sequence haplotypes identified in the course of the investigation. Only sites in the gene and 3' flanking region were considered.

Appendix D

Research Publications

A small proportion of the research presented in this thesis was published prior to submission. This appendix includes reproductions of the relevant articles:

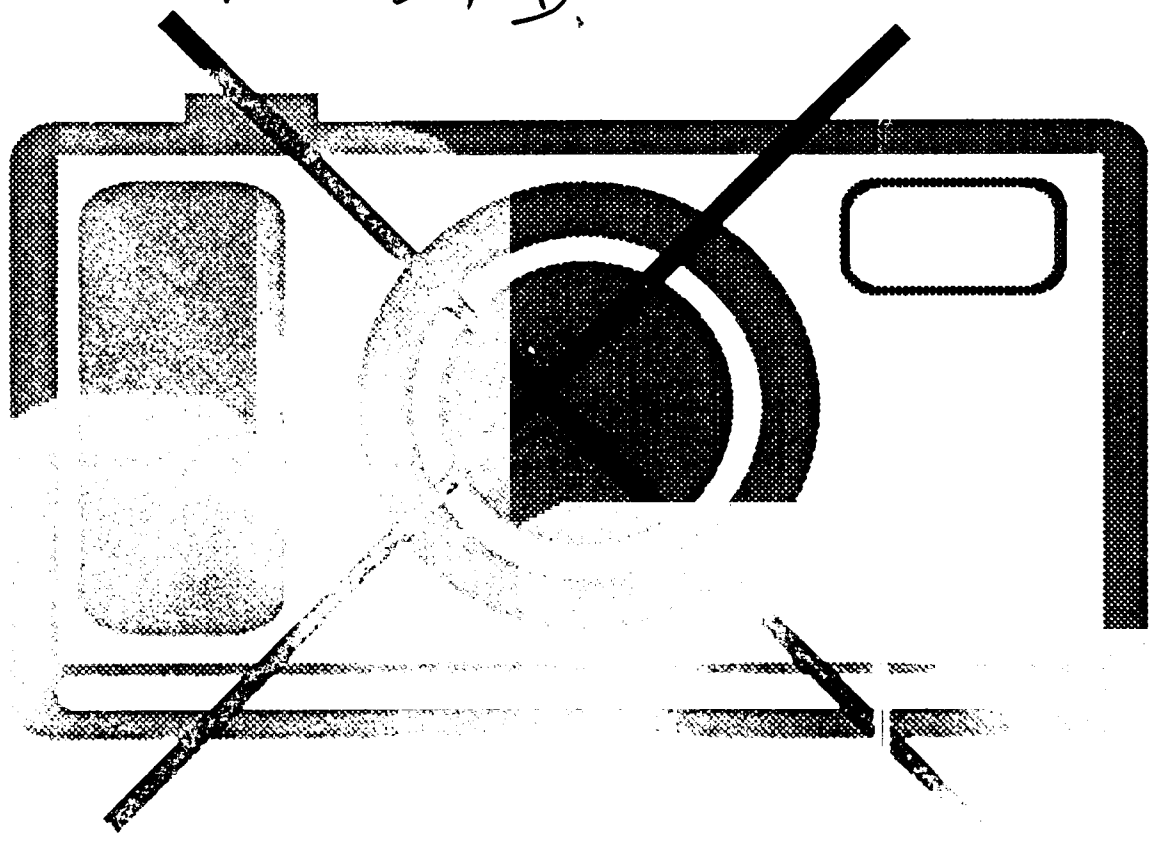
Fullerton SM, Harding RM, Boyce AJ and Clegg JB (1994) Molecular and population genetic analysis of allelic sequence diversity at the human ß-globin locus. *Proceedings of the National Academy of Sciences USA* **91**: 1805-1809.

and

Fullerton SM and Clegg JB (1994) *Hpa* I, *Hind* III, and *Bam* HI polymorphisms 3' of the human ß-globin gene can be detected by a single polymerase chain reaction amplification product. *American Journal of Hematology* **47**: 256.

PUBLISHED PAPERS NOT FILMED FOR COPYRIGHT REASONS

APPENDIX D.



Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 1805–1809, March 1994
Evolution

Molecular and population genetic analysis of allelic sequence diversity at the human β-globin locus

(Melanesia/allele-specific amplification/recombination/gene conversion/evolution)

S. M. FULLERTON*†, R. M. HARDING*, A. J. BOYCE‡, AND J. B. CLEGG*

*Medical Research Council Molecular Haematology Unit, Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford, OX3 9DU, United Kingdom; and ‡Institute of Biological Anthropology, University of Oxford, 58 Banbury Road, Oxford, OX2 6QS, United Kingdom

Communicated by David Weatherall, October 25, 1993 (received for review September 24, 1993)

ABSTRACT **Allelic sequence polymorphism at the β-globin locus was investigated in a group of 36 Melanesians. A 3-kilobase fragment containing the gene and its flanking regions was sequenced in 60 normal (β^A) and 12 thalassemic (intron 1, position 5, G → C) chromosomes. Haplotype relationships between linked polymorphisms were derived by allele-specific PCR amplification and sequencing. Seventeen nucleotide polymorphisms and 2 length variants were identified, and these sites segregated as 17 sequence haplotypes in the normal chromosomes. This haplotype diversity is higher than that expected on the basis of the nucleotide polymorphism observed and is probably due to recombination and gene conversion. Nucleotide diversity at synonymous sites in the sample is 0.14%, suggesting an average age of sequence divergence of ≈450,000 years, consistent with that expected for a neutrally evolving human nuclear locus.**

Molecular population genetic analysis of nuclear loci was pioneered by Kreitman in a study of the alcohol dehydrogenase (*Adh*) locus in *Drosophila melanogaster* (1). That survey of 11 alleles revealed extensive nucleotide polymorphism, which has since been shown to reflect the effects of both recombination (2) and balancing selection (3) on the *Adh* gene. Although allelic sequencing is the ideal approach to investigating both molecular mechanisms and population genetic questions, the stratagem of using isochromosomal species lines to simplify analysis of diploid loci is not applicable to humans. Studies of human DNA polymorphism have therefore been mostly limited to sequence analysis of haploid mitochondrial DNA (mtDNA) and to scoring of restriction fragment length polymorphisms (RFLPs) at nuclear loci (4). Allelic sequence analysis of nuclear genes has remained, largely for technical reasons, an elusive aim.

The polymerase chain reaction (PCR) (5) and accompanying innovations have made rapid sequencing of genomic DNA commonplace, but because PCR simultaneously amplifies both alleles of diploid loci, linkage relationships in compound heterozygotes are obscured. In most instances, amplified DNA must be subcloned, a time-consuming procedure with the potential for introducing sequence errors that are unacceptable in population-genetic analyses. We have used allele-specific PCR amplification (6) to circumvent these difficulties and extend the analysis of human nuclear genes by determining allelic sequence variation in a 3-kilobase (kb) region surrounding the human β-globin locus.

The β-globin locus has been one of the most intensively studied of all human loci, not least because of its association with severe inherited hemoglobinopathies such as sickle cell anemia and β-thalassemia, some of the commonest genetic diseases in the world. The earliest examples of prenatal diagnosis of these disorders using DNA analysis relied on linkage disequilibrium of β-globin gene mutations with adjacent RFLPs, over 20 of which have now been identified (7). A comprehensive set of β-globin gene cluster RFLP haplotypes has been amassed for many populations (8), and it was the availability of these data that first prompted analysis of population relationships using DNA markers (9). This well-characterized locus is thus an ideal subject for detailed analysis of allelic sequence variation in human populations.

Here, we report the results of our initial survey of nucleotide polymorphism at the β-globin locus§ in a group of Melanesians from northern Vanuatu, an island archipelago in the southwest Pacific Ocean.

MATERIALS AND METHODS

Population Sampled. Genomic DNAs, extracted from the umbilical cord bloods (10) of 36 unrelated Melanesians (24 normal individuals born on the island of Espiritu Santo in northern Vanuatu and 12 β-thalassemia heterozygotes from a neighboring island, Maewo), were used.

Amplification and Sequencing of Diploid Template. A 3.067-kb fragment containing the human β-globin gene and 1.4 kb of flanking sequence was PCR-amplified and directly sequenced (11). Genomic template (400 ng) was subjected to 35 cycles of amplification with primers corresponding to positions 10547–10572 and 13592–13613 of the human β-globin gene reference sequence (12). Ten picomoles of each primer (3′ primer biotinylated), 50 mM KCl, 10 mM Tris chloride (pH 8.4), 1.5 mM $MgCl_2$, 0.001% gelatin, 200 mM dNTPs, 1.85 μg of gene *32* protein (Pharmacia) (13), and 2 units of *Taq* polymerase were included in each 100-μl reaction. Cycle conditions were 1 min at 94°C, 2 min at 60°C, and 5 min at 72°C.

Four reactions per sample were combined, and single-stranded template was prepared by using streptavidin-coated magnetic beads (Dynal, Great Neck, NY) for capture and purification of the biotinylated strand. This template was sequenced with Sequenase T7 DNA polymerase (United States Biochemical) with eight internal sequencing primers (18–24 bases long) spaced at ≈350-base-pair (bp) intervals. Identical sequences from different individuals were run out in parallel, with corresponding dideoxy-NTP (ddNTP) termination reactions loaded in adjacent wells (Fig. 1).

Allele-Specific Amplification for Linkage Determination. To determine linkage relationships in compound heterozygotes, the amplification refractory mutation system (ARMS) (6) was used. An ARMS primer, incorporating an additional internal (position −3) mismatch (14), was used to permit *Taq* polymerase extension from only one of the two bases present at

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: RFLP, restriction fragment length polymorphism; ARMS, amplification refractory mutation system.

†To whom reprint requests should be addressed.

§The sequences reported in this paper have been deposited in the GenBank data base (accession nos. L26462–L26478).

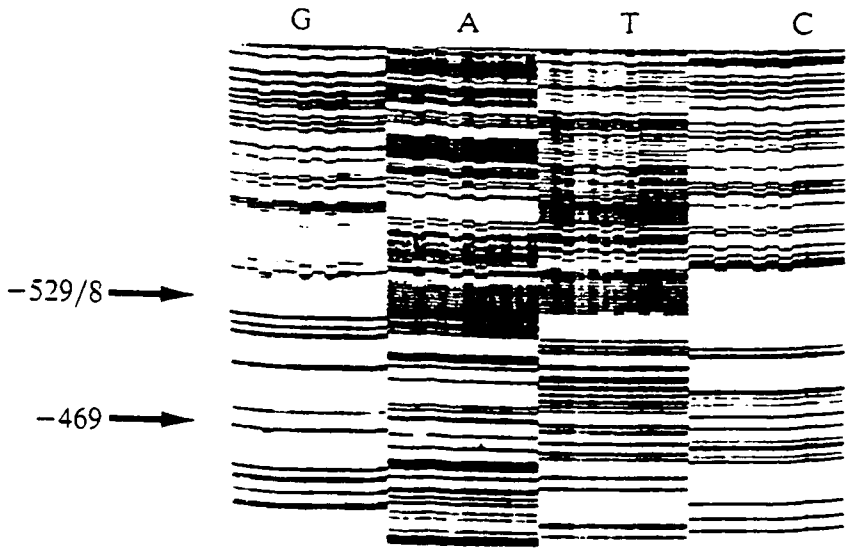


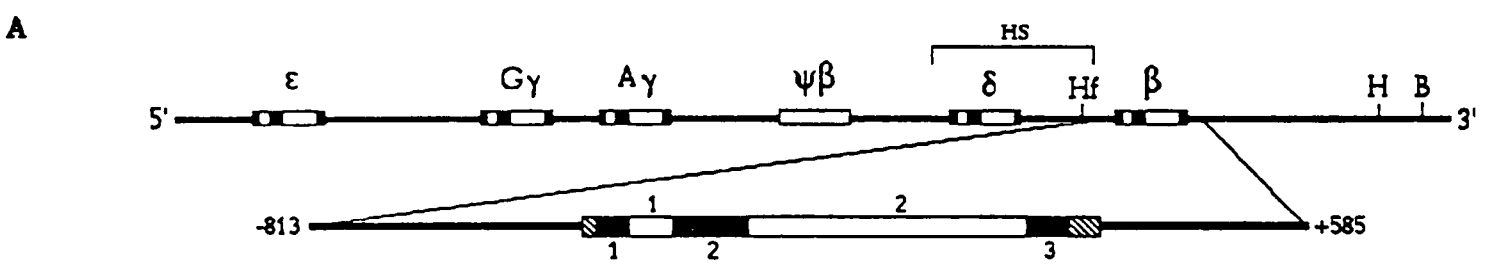
FIG. 1. DNA sequence of 12 allele-specific templates amplified by the ARMS (6) method. Termination reactions are loaded in parallel and from left to right are ddGTP (G) 1–12, ddATP (A) 1–12, ddTTP (T) 1–12, and ddCTP (C) 1–12. Nucleotides −636 to −428 of the β-globin gene cap site, 3′ → 5′ strand, are shown. The AT repeat region polymorphism (−529 and −528) is indicated, as is a downstream A → G difference in sample 3 at position −469. In heterozygous samples, the region upstream of the purine–pyrimidine repeat is uninterpretable because of length differences associated with the insertion polymorphism.

a heterozygous site. Only the sequence that lies between the preferred base and the reverse primer is amplified, resulting in an allele-specific PCR product. Sites of high heterozygosity, with proximity to either the 5′ or 3′ end of the 3.0-kb fragment (i.e., positions −703, +183, +339, +492 in Fig. 2*B*), were used in the ARMS amplifications. Allele-specific products ranging in size from 2.6 to 2.9 kb were routinely amplified to high yield and specificity, giving unambiguous DNA sequence (Fig. 1).

Genomic DNAs were amplified as described above by using the appropriate ARMS primer and complementary biotinylated 5′ or 3′ primer (annealing temperature was adjusted as required for specificity). Template preparation and sequencing was as described for diploid amplifications. With the sequence of one allele directly determined, most of the sequence of the corresponding allele was inferred by reference to the original diploid sequence. The second allele was specifically amplified and sequenced only to resolve the length polymorphisms at positions −521 and −529. In each case there were no discrepancies between the haploid and diploid sequences obtained from a given individual.

RFLP Analysis of Adjacent Sites. Amplified genomic DNA was digested with restriction enzymes to score RFLP sites outside the sequenced region. PCR primers for the *Hin*fI site 5′ to the β-globin gene were as described (17). Primers used to type the 3′ *Hin*dIII and *Bam*HI sites will be described elsewhere (S.M.F. and J.B.C., unpublished data).

Test for Population Subdivision. To evaluate subdivision, we used the single subpopulation test of Strobeck (18). This test compares two estimates of diversity, $\theta = 4N\mu$, where N is the population size and μ is the mutation rate. When assuming the infinite "alleles" model, θ is estimated from the



B

Ref	Haplotype Frequency Santo	Maewo	-989 Hinf I	-703 T Rsa I	-551 T	-541 T	-529 -	-528 T	-521 C	-469 T	-340 T	-316 T	Exon1 6 C	Intron 2 16 C Ava II	74 G	81 C	666 T	101 G	181 G	183 C	339 A Hinf I	492 G	-7000 Hind III	-8000 Bam HI
				5′ Flanking									Gene					3′ Flanking						
A1	10	2	+/-	C	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	+/-	+
A2	5	3	+/-	C	C	•	AT	A	•	•	•	•	•	•	•	•	•	•	•	•	•	•	-,+	+/-,+
A3	1		-	C	C	•	AT	A	•	•	•	C	•	•	•	•	•	•	•	•	•	•	-	+
A4	2	2	+	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	+/-	+
B1	9	2	+/-	C	C	•	AT	A	•	•	•	•	•	•	T	•	•	•	•	•	T	T	+	-
B2	2		+/-	•	C	•	AT	A	•	•	•	•	•	•	T	•	•	•	•	•	T	T	+	-
B3	1	1	+	C	C	•	AT	•	•	•	•	•	•	•	T	•	•	•	•	•	T	T	+	-
B4	1		+	C	C	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	T	T	+	-
B5	2		+	•	•	•	•	A	•	C	•	•	•	•	T	•	•	•	•	•	T	T	+	+/-
B6	1		+	•	•	•	AT	A	-(2)	•	C	•	•	•	T	•	•	•	•	•	T	T	+	-
C1	6		+/-	•	•	•	•	A	•	C	C	•	T	G	T	•	C	•	•	A	•	T	-	+
C2	1		+	•	•	C	•	A	•	C	C	•	T	G	T	•	C	•	•	A	•	T	-	+
C3	1		+	•	•	•	•	A	-(1)	•	C	•	T	G	T	•	C	•	A	A	•	T	-	+
C4	2	1	+/-	C	C	•	AT	A	•	•	C	•	T	•	T	•	C	•	•	A	•	T	-	+
D1	2	1	+/-	•	•	•	•	•	•	•	C	•	T	G	T	T	C	C	•	A	•	T	-	+
D2	1		+	•	•	•	•	•	•	•	C	•	•	G	T	T	C	C	•	A	•	T	-	+
D3	1		-	C	•	•	•	•	•	•	C	•	T	G	T	T	C	C	•	A	•	T	-	+

FIG. 2. (*A*) The 3-kb sequenced region and its relationship to the human β-globin gene cluster. The top line shows the β-globin gene cluster. RFLP sites are Hf, *Hin*fI; H, *Hin*dIII, and B, *Bam*HI. HS indicates the approximate boundaries of a previously identified 9.1-kb hotspot for recombination (32). Below the cluster, in expanded form, is the 3-kb genomic region that was sequenced. □, Introns (numbered above the boxes); ■, exons (numbered below the boxes); and ▨, untranslated regions of the β-globin gene. (*B*) Sequence variation observed in the two island samples of normal β^A chromosomes from northern Vanuatu. Variant positions are numbered at the 5′ end in relation to the β-globin gene cap site and at the 3′ end in relation to its polyadenylylation site, while polymorphisms within the gene are labeled according to functional region (i.e., intron 2: 81 is the 81st nucleotide of the second intron of the gene). Sequence haplotypes are arranged in the order of sequence similarity and labeled A, B, C, or D (see text). Stars indicate identity with the β-globin reference sequence (12) as defined on the first line, and dashes indicate site absence. At −521, (1) and (2) refer to the number of nucleotides deleted in the sequence. The RFLP sites flanking the sequenced region are shown, as are recognized RFLPs associated with observed nucleotide differences.

observed number of alleles or sequence haplotypes. When assuming the infinite "sites" model, θ is also estimated from the number of polymorphic sites (s_n) observed in a sample of n DNA sequences:

$$\theta = s_n \Big/ \sum_{i}^{n-1} (1/i).$$

This estimate of θ is used to calculate the expected number of sequence haplotypes, k:

$$k = 1 + \theta/(\theta + 1) + \ldots \theta/[\theta + (n - 1)].$$

Whereas the observed number of haplotypes should not differ significantly from the expected number when there is random mating and no intragenic recombination, population subdivision leads to fewer than the expected number of haplotypes being observed.

RESULTS

Nucleotide Polymorphism in Northern Vanuatu. We examined nucleotide polymorphism at the β-globin locus in 60 normal (β^A) chromosomes and 12 thalassemic (intron 1, position 5, G → C) chromosomes from 36 unrelated Melanesians. Seventeen dimorphic nucleotide positions were identified within the 3 kb sequenced (Fig. 2): 7 in the 5′ flanking region (814 bp total length), 1 (synonymous substitution) in the exons of the gene (444 bp), 4 in its intron sequences (980 bp), and 5 in the 3′ flanking DNA (582 bp). No polymorphisms were found in either the 5′ or 3′ untranslated regions of the gene. Ten of the 17 base changes observed were transitions, only 1 of which results from mutation at a CpG dinucleotide site (position +181). Two length polymorphisms were also observed in the sequenced region: a 2-bp AT insertion at −529 and either a 1- or 2-bp deletion at position −521. The AT dinucleotide insert accompanies either the presence or absence of a T → A replacement at −528, hence its previous report as an +ATA, −T event (19).

The dimorphisms observed at positions −541, −469, −316, +101, +181, +183, and +492 have not been reported previously to our knowledge, although the +101, +183, and +492 variants are present in Caucasians (S.M.F., unpublished data). The remaining 12 polymorphisms, including all of those within the gene, as well as the −521 and −529 length variants, have been observed in Caucasian, Asian, and African populations (16, 19–24).

There is an average of one nucleotide polymorphism every 175 bp and 1 in 700 bp vary between any two randomly chosen sequences in the Santo sample (n = 48)—i.e., nucleotide diversity at the locus, estimated from the number of observed polymorphic sites (25), is 0.14 ± 0.05%, a value not dissimilar to that estimated from RFLPs in the β-globin gene cluster (26). This level of β-globin nucleotide polymorphism is consistent with estimates of diversity at other loci. A recent survey using published cDNA and genomic sequences for 49 human nuclear genes found an average nucleotide diversity at silent sites of 0.076% (27), which is not substantially different from our own estimate of 0.14%. Although nucleotide polymorphisms are apparently nonuniformly distributed across the sequenced region, there is no significant difference in nucleotide diversity of "effectively silent" (synonymous or noncoding) sites (1) between coding, noncoding, and flanking regions sequenced (Fig. 3). However, nucleotide diversity in the introns is lower than in any other portion of the sequenced region.

Sequence Haplotype Polymorphism in Northern Vanuatu. The linkage relationships between the nucleotide polymorphisms observed in the two island populations are given in Fig. 2*B*. In the Espiritu Santo sample, the 19 variable

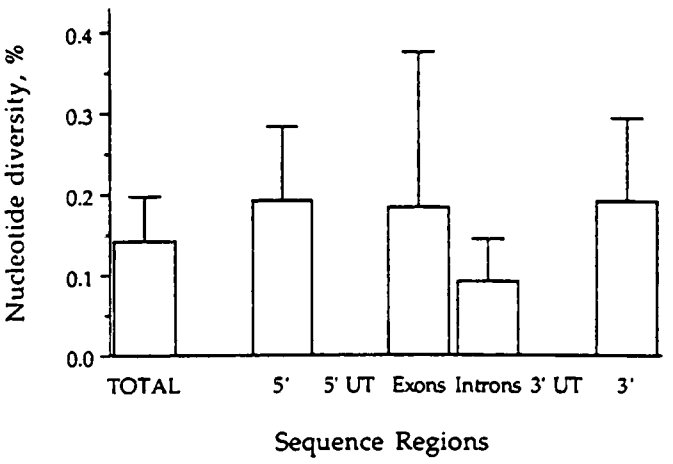


FIG. 3. Bar graph showing nucleotide diversity (25) relative to the number of effectively silent (1) sites in the 3-kb region. Nucleotide diversity within the 5′ and 3′ flanking regions and in combined coding (exons) and noncoding (introns) regions of the gene is shown to the right of the overall (total) estimate for the sequence. UT, untranslated region. Brackets represent the standard error of each estimate.

positions were found to cosegregate as 17 sequence haplotypes ranging in frequency from 2% to 21%. Seven of these haplotypes accounted for the 12 normal β^A chromosomes from Maewo. The distribution of normal haplotypes on both islands is similar, despite the different sample sizes. The β-thalassemia mutation, which is found in Vanuatu almost exclusively on Maewo (29), was present on the same sequence haplotype background (*A4*) in all 12 cases.

Haplotype polymorphism at the β-globin locus has previously been described in terms of gene "frameworks" or RFLP linkage groups. The 17 Melanesian sequence haplotypes fall into four groups, designated A (42% of the β^A chromosomes), B (32%), C (18%), and D (8%), whose within-gene polymorphisms correspond respectively to the four previously described β-globin gene frameworks 1, 2, 3-Asian, and 3 (20). The same haplotype groups are identified independently by unweighted pair group method with arithmetic mean (UPGMA) analysis (data not shown). Each group is distinguished by a characteristic pattern of linked polymorphic sites in the gene and its 3′ flanking region.

Our survey is unusual in observing four haplotype groups in one population. Other studies have only ever identified three gene frameworks, hence groups, in a particular ethnic sample: A, B, and D in Mediterraneans and A, B, and C in Asian Indians, Southeast Asians, Chinese, and American Blacks (7). Group D (framework 3) gene sequences, which have previously only been observed in Caucasian populations, make up the smallest proportion of the Melanesian sequence haplotypes. However, earlier analyses, which largely relied on RFLP analysis for identification of gene frameworks, could not have distinguished C and D type sequences, so the presence of group D β-globin gene sequence haplotypes in Asian populations may have been inadvertently overlooked.

Fig. 2 also shows the relationship of the 17 sequence haplotypes to surrounding restriction enzyme sites that make up the 3′ β-globin gene cluster RFLP haplotype. The *Hin*dIII and *Bam*HI sites that lie, respectively, 7 and 8 kb 3′ to the gene, form four haplotypes: ++, −−, −+, and +−. The C and D sequence groups are together associated with the −+ RFLP haplotype. Another pattern, +−, exclusively defines B-type sequences, although *B5* is also associated with the ++ haplotype. However, A-type sequence haplotypes, *A2* in particular, are found with three RFLP haplotypes, −+, ++, and −−. The *Hin*fI polymorphism, although lying only 175 bp 5′ of the start of the 3-kb region, is randomly associated with sequences from each haplotype group, as noted previously (30). Clearly, while RFLP analysis tends to confirm major distinctions observed at the nucleotide level, sequence anal-

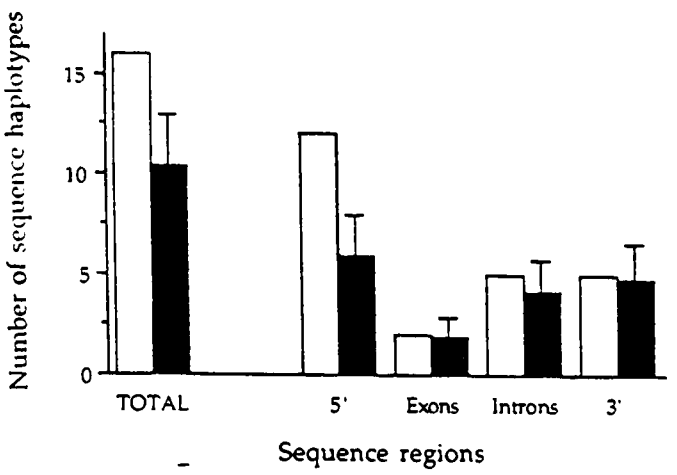


FIG. 4. Bar graph showing observed (□) versus expected (■) haplotype diversity in the Espiritu Santo sample as calculated by the Strobeck test (18). Results for the total sequence are given in the left-most comparison. Similar comparisons, calculated separately for the 5′ and 3′ flanking DNA and for combined coding (exons) and noncoding (introns) regions of the gene, are also shown. Brackets represent the standard error of each estimate. Length polymorphism was not taken into account in these comparisons.

ysis reveals chromosomal relationships that would have been overlooked or misinterpreted if only RFLP analysis had been used.

5′ RFLP haplotypes found in this study are similar to those identified in a previous sample from Vanuatu (29) and appear to be in linkage equilibrium with the 3′ haplotypes (data not shown).

Comparison of Nucleotide and Sequence Haplotype Polymorphism. We looked for evidence of population subdivision by comparing the observed number of sequence haplotypes with the number expected given the nucleotide diversity (18) (Fig. 4). However, rather than observing a reduction in haplotype diversity consistent with population subdivision, we found significantly more sequence haplotypes in Espiritu Santo than expected. The smaller Maewo sample was not analyzed. When similar comparisons are made for particular regions of the sequence (Fig. 4), haplotype diversity is only excessive in the 5′ flanking DNA. This nonuniform relationship between the nucleotide and haplotype polymorphism observed is suggestive of a molecular rather than demographic process; indeed, excess haplotype diversity is expected at loci undergoing recombination (31).

DISCUSSION

We used DNA sequence analysis to determine nucleotide polymorphism at the human β-globin locus in a Melanesian population. Although nucleotide diversity in Vanuatu is consistent with global estimates derived from cDNA and RFLP studies, the remarkable feature of the data is the considerable haplotype variation revealed by sequencing. Not surprisingly, previous approaches gave no intimation of this complexity, primarily because excess haplotype diversity, rather than resulting from the accumulation of new allelic variants via point mutation, is probably due to the shuffling of existing nucleotide polymorphisms by recombination, a scenario first envisaged by Watt (28) and later developed theoretically by Strobeck and Morgan (31).

The conclusion that recombination in the 5′ flanking region has generated much of the observed haplotype diversity is supported by the distribution of nucleotide sites among the sequence haplotypes. Haplotype groups, which most likely represent distinct sequence lineages, are each polymorphic for the same sites in the region 5′ to the β-globin gene. The likelihood of so many nucleotide dimorphisms recurring in different lineages as a consequence of random mutation is very small; they are better explained as resulting from interallelic recombination and gene conversion. Similar "recurrences" affecting sites that lie within the β-globin gene are present in sequences *C4* and *D2*. However, other differences within groups, such as those defining *A3* (−316), *C2* (−541), and *C3* (+181), are better ascribed to new substitutions arising on existing sequence haplotype backgrounds.

The excess diversity attributable to recombination is found predominantly in the 5′ flanking DNA of our 3-kb region. This is almost certainly not the full extent of recombinatorial activity, however, as the 5′ *Hin*fI site is also randomly associated with the haplotype groups defined by sequencing. Conceivably, then, the region immediately 5′ to the β-globin gene may be strongly influenced by, if not actually an extension of, an upstream recombination "hotspot" previously defined by linkage analysis (32) (see Fig. 2*A*). Recombination rates for this hotspot have been estimated to be 3–30 times higher than those of surrounding regions (33). Although probably initiating recombination outside the 3-kb sequenced region, the hotspot could be responsible for the recombination-driven haplotype diversity we have observed. Certainly, there are examples in both cultured mammalian cells and yeast of homologous recombination and gene conversion occurring at considerable distances from a recombinogenic origin (refs. 34 and 35 for example).

Although most of the interallelic exchanges occur in DNA 5′ of the β-globin gene, there are exceptions, such as sequence haplotypes *C4* and *D2*, which may have resulted from short-tract gene conversion around single polymorphic sites within the gene itself. Similar examples of possible conversion-generated diversity may be found among common hemoglobinopathies such as HbS, HbE, and some β-thalassemias. Many of these mutations are found on diverse 3′ RFLP haplotype backgrounds despite evidence that they have only recently reached high frequencies by malarial selection (8). While this has been ascribed to multiple independent mutations, interallelic gene conversion of single original mutations seems a more plausible explanation, especially in view of the fact that most of these putative conversions are found in the 5′-most (recombinogenic) part of the gene (36).

At the 3′ end of the sequenced region, where recombination and gene conversion have not been acting to increase sequence haplotype diversity, there is no evidence from the Strobeck test for increased genetic drift induced by population substructure or expansion from a recent population bottleneck. All of the previously identified β-globin gene frameworks are present in our northern Vanuatu sample, and the majority of the polymorphisms observed are known to be globally distributed. Such variation belies the archipelago's remote Pacific location and its relatively recent [within the past 3500 years (37)] settlement history. In contrast, limited evidence for lower average heterozygosity of both mtDNA loci (38) and RFLP sites (15), suggestive of increased genetic drift, has been obtained in studies that have compared Melanesians (from Papua New Guinea and the Solomon Islands) with other ethnic groups.

The number of polymorphic sites in our sample from northern Vanuatu provides a recombination-independent estimate of the age of the observed variation. As 1.6% difference (39) has accumulated at the β-globin locus during the 5 million years (40) separating chimpanzees from humans, 0.14% nucleotide diversity corresponds to an average age of sequence divergence of ≈450,000 years. Similarly, the pairwise difference between *A3* and *D1*, the two most divergent haplotypes in the sample, suggests that the 3′ nonrecombining portions of these sequences may have persisted for more than twice that time, ≈1.1 million years. These ages, based on data from a single population, are consistent with the coalescence time expected for a neutrally evolving human nuclear locus, given a long-term effective population size of 10,000 (41).

Therefore, our data cannot support the suggestion that the human species underwent a population bottleneck 200,000 years ago, as is controversially suggested by mtDNA analysis (42). The discrepancy between these estimated ages is likely to reflect the 4-fold greater impact of genetic drift, which proportionately reduces coalescence time for mtDNA compared with neutral nuclear DNA (41). To explore the genetic contribution to the modern gene pool of regional populations prior to 200,000 years ago, then, requires investigation of human nuclear loci.

Although only one locale has been sampled, two important conclusions have been drawn from our data. First, a substantial part of the haplotype diversity observed has been generated by recombination and gene conversion. Second, observed nucleotide diversity suggests an average age of sequence divergence of at least 450,000 years, supporting coalescence times predicted from neutral theory. This study of allelic variation at the human β-globin locus thus demonstrates the value of sequence analysis for the investigation of molecular and population processes. A fuller appreciation of the nature and extent of β-globin nucleotide polymorphism awaits similar analyses of other human populations. Global investigation of β-globin diversity alone will not solve the puzzle of human origins, however, and many other loci will need to be studied before consensus regarding the evolutionary history of *Homo sapiens* is realized.

We thank Dr. Mary Ganczakowski for providing the samples used in our analysis and Drs. Rhona Borts, Peter Donnelly, Jonathan Flint, Adrian Hill, Ed Louis, and an anonymous reviewer for helpful comments on the manuscript. S.M.F. was supported by grants from the Rhodes and Wellcome Trusts.

1. Kreitman, M. (1983) *Nature (London)* **304**, 412–417.
2. Stephens, J. C. & Nei, M. (1985) *J. Mol. Evol.* **22**, 289–300.
3. Hudson, R. R., Kreitman, M. & Aguade, M. (1987) *Genetics* **116**, 153–159.
4. Devor, E. J., ed. (1992) *Molecular Applications in Biological Anthropology* (Cambridge Univ. Press, Cambridge, U.K.).
5. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1988) *Science* **239**, 487–491.
6. Newton, C. R., Graham, A., Heptinstall, L. E., Powell, S. J., Summers, C., Kalsheker, N., Smith, J. C. & Markham, A. F. (1989) *Nucleic Acids Res.* **17**, 2503–2516.
7. Antonarakis, S. E., Kazazian, H. H., Jr., & Orkin, S. H. (1985) *Hum. Genet.* **69**, 1–14.
8. Flint, J., Harding, R. M., Clegg, J. B. & Boyce, A. J. (1993) *Hum. Genet.* **91**, 91–117.
9. Wainscoat, J. S., Hill, A. V. S., Boyce, A. J., Flint, J., Hernandez, M., Thein, S. L., Old, J. M., Lynch, J. R., Falusi, A. G., Weatherall, D. J. & Clegg, J. B. (1986) *Nature (London)* **319**, 491–493.
10. Old, J. M. & Higgs, D. R. (1983) in *The Thalassemias: Methods in Hematology*, ed. Weatherall, D. J. (Churchill Livingstone, London), pp. 74–102.
11. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
12. Poncz, M., Schwartz, E., Ballantine, M. & Surrey, S. (1983) *J. Biol. Chem.* **258**, 11599–11609.
13. Schwarz, K., Hansen-Hagge, T. & Bartram, C. (1990) *Nucleic Acids Res.* **18**, 1079.
14. Kwok, S., Kellogg, D. E., McKinney, N., Spasic, D., Goda, L., Levenson, C. & Sninsky, J. J. (1990) *Nucleic Acids Res.* **18**, 999–1005.
15. Bowcock, A. M., Kidd, J. R., Mountain, J. L., Hebert, J. M., Carotenuto, L., Kidd, K. K. & Cavalli-Sforza, L. L. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 839–843.
16. Moschonas, N., de Boer, E. & Flavell, R. A. (1982) *Nucleic Acids Res.* **10**, 2109–2120.
17. Sutton, M., Bouhassira, E. E. & Nagel, R. L. (1989) *Am. J. Hematol.* **32**, 66–69.
18. Strobeck, C. (1987) *Genetics* **117**, 149–155.
19. Semenza, G. L., Malladi, P., Surrey, S., Delgrosso, K., Poncz, M. & Schwartz, E. (1984) *J. Biol. Chem.* **259**, 6045–6048.
20. Orkin, S. H., Kazazian, H. H., Jr., Antonarakis, S. E., Goff, S. C., Boehm, C. D., Sexton, J. P., Waber, P. G. & Giardina, P. J. V. (1982) *Nature (London)* **296**, 627–631.
21. Semenza, G. L., Delgrosso, K., Poncz, M., Malladi, P., Schwartz, E. & Surrey, S. (1984) *Cell* **39**, 123–128.
22. Cheboune, Y., Pagnier, J., Trabuchet, G., Faure, C., Verdier, G., Labie, D. & Nigon, V. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 4431–4435.
23. Trabuchet, G., Elion, J., Baudot, G., Pagnier, J., Bouhass, R., Nigon, V. M., Labie, D. & Krishnamoorthy, R. (1991) *Hum. Biol.* **63**, 241–252.
24. Lapouméroulie, C., Dunda, O., Ducrocq, R., Trabuchet, G., Mony-Lobé, M., Bodo, J. M., Carnevale, P., Labie, D., Elion, J. & Krishnamoorthy, R. (1992) *Hum. Genet.* **89**, 333–337.
25. Watterson, G. A. (1975) *Theor. Popul. Biol.* **7**, 256–276.
26. Kazazian, H. H., Jr., Chakravarti, A., Orkin, S. H. & Antonarakis, S. E. (1983) in *Evolution of Genes and Proteins*, eds. Nei, M. & Koehn, R. K. (Sinauer, Sunderland, MA), pp. 137–146.
27. Li, W.-H. & Sadler, L. A. (1991) *Genetics* **129**, 513–523.
28. Watt, W. B. (1972) *Am. Nat.* **106**, 737–753.
29. Hill, A. V. S., Bowden, D. K., O'Shaughnessy, D. F., Weatherall, D. J. & Clegg, J. B. (1988) *Blood* **72**, 9–14.
30. Kohen, G., Philippe, N. & Godet, J. (1982) *Hum. Genet.* **62**, 121–123.
31. Strobeck, C. & Morgan, K. (1978) *Genetics* **88**, 829–844.
32. Antonarakis, S. E., Boehm, C. D., Giardina, P. J. V. & Kazazian, H. H., Jr. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 137–141.
33. Chakravarti, A., Buetow, K. H., Antonarakis, S. E., Waber, P. G., Boehm, C. D. & Kazazian, H. H., Jr. (1984) *Am. J. Hum. Genet.* **36**, 1239–1258.
34. Wahls, W. P., Wallace, L. J. & Moore, P. D. (1990) *Mol. Cell. Biol.* **10**, 785–793.
35. Symington, L. S., Brown, A., Oliver, S. G., Greenwell, P. & Petes, T. D. (1991) *Genetics* **128**, 717–727.
36. Matsuno, Y., Yamashiro, Y., Yamamoto, K., Hattori, Y., Yamamoto, K., Ohba, Y. & Miyaji, T. (1992) *Hum. Genet.* **88**, 357–358.
37. Bellwood, P. (1989) in *The Colonization of the Pacific: A Genetic Trail*, eds. Hill, A. V. S. & Serjeantson, S. W. (Oxford Univ. Press, Oxford), pp. 1–59.
38. Stoneking, M. & Wilson, A. C. (1989) in *The Colonization of the Pacific: A Genetic Trail*, eds. Hill, A. V. S. & Serjeantson, S. W. (Oxford Univ. Press, Oxford), pp. 215–245.
39. Savatier, P., Trabuchet, G., Faure, C., Cheboune, Y., Gouy, M., Verdier, G. & Nigon, V. M. (1985) *J. Mol. Biol.* **182**, 21–29.
40. Horai, S., Satta, Y., Hayasaka, K., Kondo, R., Inoue, T., Ishida, T., Hayashi, S. & Takahata, N. (1992) *J. Mol. Evol.* **35**, 32–43.
41. Takahata, N. (1993) *Mol. Biol. Evol.* **10**, 2–22.
42. Cann, R. L., Stoneking, M. & Wilson, A. C. (1987) *Nature (London)* **325**, 31–36.

unresponsive idiopathic immune thrombocytopenic purpura after short-course therapy with recombinant alpha 2b interferon. Blood 74:1894–1897, 1989.
3. Matthey F, Ardeman S, Jones L, Newland AC: Bleeding in immune thrombocytopenic purpura after alpha-interferon. Lancet i:471–472, 1990.
4. Stellini R, Rossi G, Paraninfo G: Interferon therapy in intravenous drug users with HIV-associated idiopathic thrombocytopenic purpura. Haematologica 77:418–420, 1992.
5. Conlon KC, Urba WJ, Smith JW, Steis RG, Longo DL, Clark JW: Exacerbation of symptoms of autoimmune disease in patients receiving alpha-interferon therapy. Cancer 65:2237–2242, 1990.

HpaI, HindIII, and BamHI Polymorphisms 3′ of the Human β-Globin Gene Can Be Detected by a Single Polymerase Chain Reaction Amplification Product

To the Editor: Recent development of a PCR-based approach to β-globin RFLP (restriction fragment length polymorphism) analysis has permitted rapid typing of the major polymorphic sites found within the human β-globin cluster [1]. Nevertheless, three commonly investigated restriction enzyme sites, the HpaI, HindIII, and BamHI sites, which lie respectively 6, 7, and 8 kb 3′ of the β-globin gene, have eluded analysis by PCR. Primers chosen to amplify the HpaI site were found to lack specificity [2] and no primers have been reported which allow RFLP analysis of the HindIII or BamHI sites. A polymorphic HinfI site 340 bp 3′ of the β-globin gene, which can be PCR-amplified [3], is a suitable replacement for the purposes of prenatal diagnosis. In order to extend analysis beyond the immediate region of the β-globin gene, however, it would also be desirable to investigate the status of the more distant RFLP sites via PCR.

The lack of specificity observed in the case of the HpaI PCR amplification product has been attributed to the location of the RFLP site within a 6.2 kb KpnI family (L1) repetitive element [2]. Misannealing of the primers to KpnI repetitive sequences in non-β-globin regions of the genome results in a heterogeneous PCR product, which does not always reflect accurately the status of the RFLP site 3′ of the β-globin gene. As the polymorphic HindIII and BamHI sites also fall within the same KpnI repeat, all three sites are subject to the same misamplification effect.

To circumvent this problem of non-specificity, we chose PCR primers that would only amplify sites within the KpnI family member 3′ of the human β-globin gene. While the 5′ primer (CAT GCC AAA TTG TAA ACG ACC ATC AAG G) anneals to KpnI sequence, the 3′ primer (GAT AAT GGC TCT ACG GAT GTG TGA GAT C) is complementary to single-copy DNA that lies outside of, but immediately adjacent to, the repetitive region [4] (see Fig. 1). The resulting 4.3 kb fragment, which can be readily produced using gp32 [5] (35 cycles of 1 min at 94 °C, 2 min at 65 °C, 5 min at 72 °C), amplifies the HpaI as well as HindIII and BamHI polymorphic sites and permits unambiguous typing of all three RFLPs, as shown. When cut, the 4,259 bp product yields fragment sizes of 3,700 and 559 bp for HpaI, 2,517 and 1,742 bp for HindIII, and 2,920 and 1,339 bp for BamHI.

The consistently heterozygous (+/−) pattern observed for the original HpaI PCR primers [2] is not observed using these primers, and results

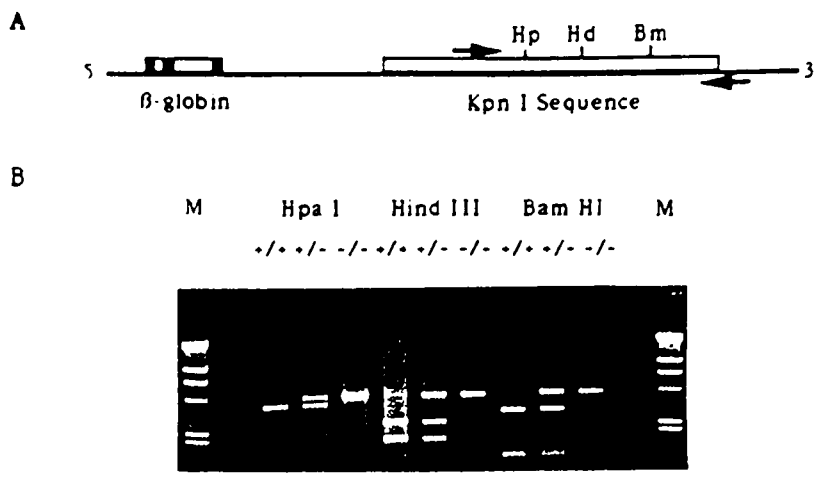


Fig. 1. A: Location of the polymorphic HpaI (Hp), HindIII (Hd), and BamHI (Bm) restriction enzyme sites, relative to the human β-globin gene and its KpnI repetitive sequence. Arrows represent the PCR amplification primers. B: Results of restriction enzyme digestion of the 4.3 kb PCR product with either HpaI, HindIII, or BamHI (1.0% agarose gel). For each enzyme, the pattern of fragments obtained when the site tested is either homozygous (+/+ or −/−) or heterozygous (+/−) is shown. Fragment sizes are given in the text. The marker (M) is 500 ng λ DNA digested with HindIII.

obtained with the 4.2 kb fragment have been confirmed by Southern blotting and probing of equivalent samples (data not shown). Our strategy demonstrates that even long repetitive sequences can be amplified from specific genomic regions if one of the two primers anneals to single copy DNA. The resulting PCR product provides a rapid and reliable means of typing the three 3′ RFLP sites and complements the protocol already in use for β-globin haplotyping by PCR.

Stephanie M. Fullerton
John B. Clegg
MRC Molecular Haematology Unit, Institute of Molecular Medicine, University of Oxford, Oxford, United Kingdom

References

1. Sutton M, Bouhassira EE, Nagel RL: Polymerase chain reaction amplification applied to the determination of β-like globin gene cluster haplotypes. Am J Hematol 32:66, 1989.
2. Wang X, Bouhassira EE: HpaI polymorphic site 3′ of the human β-globin gene is inside a repetitive sequence and cannot be ascertained by polymerase chain reaction. Am J Hematol 39:226, 1992.
3. Semenza GL, Dowling CE, Kazazian HH: HinfI polymorphism 3′ to the human β-globin gene detected by the polymerase chain reaction (PCR). Nucleic Acids Res 17:2376, 1989.
4. Hattori M, Hidaka S, Sakaki Y: Sequence analysis of a KpnI family member near the 3′ end of human β-globin gene. Nucleic Acids Res 13:7813, 1985.
5. Schwarz K, Hansen-Hagge T, Bartram C: Improved yields of long PCR products using gene 32 protein. Nucleic Acids Res 18:1079, 1990.

REFERENCES

References

Antonarakis SE, Boehm CD, Giardina PJV and Kazazian Jr. HH (1982) Nonrandom association of polymorphic restriction sites in the ß-globin gene cluster. *Proceedings of the National Academy of Sciences USA* **79**: 137-141.

Antonarakis SE, Boehm CD, Serjeant GR, Theisen CE, Dover GJ and Kazazian Jr. HH (1984) Origin of the $ß^S$-globin gene in Blacks: the contribution of recurrent mutation or gene conversion or both. *Proceedings of the National Academy of Sciences USA* **81**: 853-856.

Antonarakis SE, Kazazian Jr. HH and Orkin SH (1985) DNA polymorphism and molecular pathology of the human globin gene clusters. *Human Genetics* **69**: 1-14.

Atcheson CL and Esposito RE (1993) Meiotic recombination in yeast. *Current Opinion in Genetics and Development* **3**: 736-744.

Ayala FJ, Escalante A, O'hUigin C and Klein J (1994) Molecular genetics of speciation and human origins. *Proceedings of the National Academy of Sciences USA* **91**: 6787-6794.

Begun DJ and Aquadro CF (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster. Nature (London)* **356**: 519-520.

Begun DJ and Aquadro CF (1993) African and North American populations of *Drosophila melanogaster* are very different at the DNA level. *Nature (London)* **365**: 548-550.

Begun DJ and Aquadro CF (1994) Evolutionary inferences from DNA variation at the 6-phosphogluconate dehydrogenase locus in natural populations of Drosophila: selection and geographic differentiation. *Genetics* **136**: 155-171.

Bellwood P (1978) *Man's conquest of the Pacific: the prehistory of Southeast Asia and Oceania.* Auckland: William Collins Publishers, Ltd.

Bellwood P (1989) The colonization of the Pacific: some current hypotheses, pp. 1-59. In Hill AVS and Serjeantson SW (eds) *The colonization of the Pacific: a genetic trail.* Oxford: Oxford University Press.

Bellwood P (1991) The Austronesian dispersal and the origin of languages. *Scientific American* **265**: 70-75.

Bergemann AD and Johnson EM (1992) The HeLa pur factor binds single-stranded DNA at a specific element conserved in gene flanking regions and origins of DNA replication. *Molecular and Cellular Biology* **12**: 1257-1265.

Bilofsky HS and Burks C (1988) The GenBank genetic sequence data bank. *Nucleic Acids Research* **16**: 1861-1864.

Blaho JA and Wells RD (1989) Left-handed Z-DNA and genetic recombination. *Progress in Nucleic Acid Research and Molecular Biology* **37**: 107-126.

Blake RD, Hess ST and Nicholson-Tuell J (1992) The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *Journal of Molecular Evolution* **34**: 189-200.

Bowcock AM, Bucci C, Hebert JM, Kidd JR, Kidd KK, Friedlaender JS and Cavalli-Sforza LL (1987) Study of 47 DNA markers in five populations from four continents. *Gene Geography* **1**: 47-64.

Bowcock AM, Hebert JM, Mountain JL, Kidd JR, Rogers J, Kidd KK and Cavalli-Sforza LL (1991a) Study of an additional 58 DNA markers in five human populations from four continents. *Gene Geography* **5**: 151-173.

Bowcock AM, Kidd JR, Mountain JL, Hebert JM, Carotenuto L, Kidd KK and Cavalli-Sforza LL (1991b) Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. *Proceedings of the National Academy of Sciences USA* **88**: 839-843.

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR and Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature (London)* **368**: 455-457.

Cai S-P, Eng B, Francombe WH, Olivieri NF, Kendall AG, Waye JS and Chui DHK (1992) Two novel ß-thalassemia mutations in the 5' and 3' noncoding regions of the ß-globin gene. *Blood* **79**: 1342-1346.

Camaschella C, Serra A, Saglio G, Tiziana Bertero M, Mazza U, Terzoli S, Brambati B, Cremonesi L, Travi M and Ferrari M (1988) Meiotic recombination in the ß globin gene cluster causing an error in prenatal diagnosis of ß thalassaemia. *Journal of Medical Genetics* **25**: 307-310.

Cann RL, Stoneking M and Wilson AC (1987) Mitochondrial DNA and human evolution. *Nature (London)* **325**: 31-36.

Cavalli-Sforza LL, Piazza A, Menozzi P and Mountain JL (1988) Reconstruction of human evolution: bringing together genetic, archaeological and linguistic data. *Proceedings of the National Academy of Sciences USA* **85**: 6002-6006.

Chakravarti A, Buetow KH, Antonarakis SE, Waber PG, Boehm CD and Kazazian Jr. HH (1984) Nonuniform recombination within the human ß-globin gene cluster. *American Journal of Human Genetics* **36**: 1239-1258.

Charlesworth B (1992) New genes sweep clean. *Nature* **356**: 475-476.

Charlesworth B, Morgan MT and Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289-1303.

Chebloune Y, Pagnier J, Trabuchet G, Faure C, Verdier G, Labie D and Nigon V (1988) Structural analysis of the 5' flanking region of the ß-globin gene in African sickle cell anemia patients: further evidence for three origins of the sickle cell mutation in Africa. *Proceedings of the National Academy of Sciences USA* **85**: 4431-4435.

Chen LZ, Easteal S, Board PG and Kirk RL (1990) Evolution of ß-globin haplotypes in human populations. *Molecular Biology and Evolution* **7**: 423-437.

Cheng T-C, Orkin SH, Antonarakis SE, Potter MJ, Sexton JP, Giardina PJV, Li A and Kazazian Jr. HH (1984) ß-thalassemia in Chinese: use of in vivo RNA analysis and oligonucleotide hybridization in systematic characterization of molecular defects. *Proceedings of the National Academy of Sciences USA* **81**: 2821-2826.

Clark AG and Whittam TS (1992) Sequencing errors and molecular evolutionary analysis. *Molecular Biology and Evolution* **9**: 744-752.

Cotton RGH (1993) Current methods of mutation detection. *Mutation Research* **285**: 125-144.

Craxton M (1991) Linear amplification sequencing, a powerful method for sequencing DNA. *Methods: Companion to Methods in Enzymology* **3**: 20-25.

Crowe JS, Cooper HJ, Smith MA, Sims MJ, Parker D and Gewert D (1991) Improved cloning efficiency of polymerase chain reaction (PCR) products after proteinase K digestion. *Nucleic Acids Research* **19**: 184.

Desmond A and Moore J (1991) *Darwin.* London: Penguin Books Ltd.

Di Rienzo A and Wilson AC (1991) Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proceedings of the National Academy of Sciences USA* **88**: 1597-1601.

Diamond J (1988) Express train to Polynesia (Archaeology, News and Views). *Nature (London)* **336**: 307-308.

Dianzani I, Camaschella C, Ponzone A and Cotton RGH (1993) Dilemmas and progress in mutation detection. *Trends in Genetics* **9**: 403-405.

Edwards AWF (1971) Distance between populations on the basis of gene frequencies. *Biometrics* **27**: 873-881.

Ellis N, Taylor A, Bengtsson BO, Kidd J, Rogers J and Goodfellow P (1990) Population structure of the human pseudoautosomal boundary. *Nature (London)* **344**: 663-665.

Engelke DR, Hoener PA and Collins FS (1988) Direct sequencing of enzymatically amplified human genomic DNA. *Proceedings of the National Academy of Sciences USA* **85**: 544-548.

Ewens WJ (1972) The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**: 87-112.

Excoffier L (1990) Evolution of human mitochondrial DNA: evidence for departure from a pure neutral model of populations at equilibrium. *Journal of Molecular Evolution* **30**: 125-139.

Excoffier L and Langaney A (1989) Origin and differentiation of human mitochondrial DNA. *American Journal of Human Genetics* **44**: 73-85.

Flint J, Harding RM, Boyce AJ and Clegg JB (1993a) The population genetics of the haemoglobinopathies. *Bailliére's Clinical Haematology* **6**: 215-262.

Flint J, Harding RM, Clegg JB and Boyce AJ (1993b) Why are some genetic diseases common?: Distinguishing selection from other processes by molecular analysis of globin gene variants. *Human Genetics* **91**: 91-117.

Frayer DW, Wolpoff MH, Thorne AG, Smith FH and Pope GG (1994) Getting it straight. *American Anthropologist* **96**: 424-438.

Fullerton SM and Clegg JB (1994) *Hpa* I, *Hin*d III, and *Bam* HI polymorphisms 3' of the human ß-globin gene can be detected by a single polymerase chain reaction amplification product. *American Journal of Hematology* **47**: 256.

Fullerton SM, Harding RM, Boyce AJ and Clegg JB (1994) Molecular and population genetic analysis of allelic sequence diversity at the human ß-globin locus. *Proceedings of the National Academy of Sciences USA* **91**: 1805-1809.

Gal S and Hohn B (1990) Direct sequencing of double-stranded DNA PCR products via removing the complementary strand with single-stranded DNA of an M13 clone. *Nucleic Acids Research* **18**: 1076.

Ganguly A and Prockop DJ (1990) Detection of single-base mutations by reaction of DNA heteroduplexes with a water-soluble carbodiimide followed by primer extension: application to products from the polymerase chain reaction. *Nucleic Acids Research* **18**: 3933-3939.

Gerhard DS, Kidd KK, Kidd JR, Egeland JA and Housman DE (1984) Identification of a recent recombination event within the human ß-globin gene cluster. *Proceedings of the National Academy of Sciences USA* **81**: 7875-7879.

Gibbons A (1992) Mitochondrial Eve: wounded, but not dead yet. *Science* **257**: 873-875.

Gibbons A (1994) Rewriting - and redating - prehistory. *Science* **263**: 1087-1088.

Giles RE, Blanc H, Cann HM and Wallace DC (1980) Maternal inheritance of human mitochondrial DNA. *Proceedings of the National Academy of Sciences USA* **77**: 6715-6719.

Grimm C, Bähler J and Kohli J (1994) *M26* recombinational hotspot and physical conversion tract analysis in the *ade6* gene of *Schizosaccharomyces pombe*. *Genetics* **135**: 41-51.

Grompe M (1993) The rapid detection of unknown mutations in nucleic acids. *Nature Genetics* **5**: 111-117.

Gyllensten UB and Allen M (1993) Sequencing of *in vitro* amplified DNA. *Methods in Enzymology* **218**: 3-16.

Gyllensten UB and Erlich HA (1988) Generation of single-stranded DNA by the polymerase chain reaction and its application to direct sequencing of the HLA-DQA locus. *Proceedings of the National Academy of Sciences USA* **85**: 7652-7656.

Hall GW, Sampietro M, Barnetson R, Fitzgerald J, McCann S and Thein SL (1993) Meiotic recombination in an Irish family with beta-thalassaemia. *Human Genetics* **92**: 28-32.

Harpending HC, Sherry ST, Rogers AR and Stoneking M (1993) The genetic structure of ancient human populations. *Current Anthropology* **34**: 483-496.

Hayashi K (1991) PCR-SSCP: a simple and sensitive method for detection of mutations in the genomic DNA. *PCR Methods and Applications* **1**: 34-38.

Hayashi K, Orita M, Suzuki Y and Sekiya T (1989) Use of labeled primers in polymerase chain reaction (LP-PCR) for a rapid detection of the product. *Nucleic Acids Research* **17**: 3605.

Hedges SB, Kumar S, Tamura K and Stoneking M (1992) Human origins and analysis of mitochondrial DNA sequences - reply. *Science* **255**: 737-739.

Hertzberg MS, Mickleson KNP and Trent RJ (1988) α-globin gene haplotypes in Polynesians: their relationships to population groups and gene rearrangements. *American Journal of Human Genetics* **43**: 971-977.

Higgs DR, Wainscoat JS, Flint J, Hill AVS, Thein SL, Nicholls RD, Teal H, Ayyub H, A. PTE, Falusi Y, Jarman A, Clegg JB and Weatherall DJ (1986) Analysis of the human α-globin cluster reveals a highly informative genetic locus. *Proceedings of the National Academy of Sciences USA* **83**: 5165-5169.

Higuchi RG and Ochman H (1989) Production of single-stranded DNA templates by exonuclease digestion following the polymerase chain reaction. *Nucleic Acids Research* **17**: 5865.

Hill A and Ward S (1988) Origin of the Hominidae: the record of African large hominoid evolution between 14 my and 4 my. *Yearbook of Physical Anthropology* **31**: 49-83.

Hill AVS, Bowden DK, O'Shaughnessy DF, Weatherall DJ and Clegg JB (1988) ß thalassemia in Melanesia: association with malaria and characterization of a common variant. *Blood* **72**: 9-14.

Hill AVS, O'Shaughnessy DF and Clegg JB (1989) Haemoglobin and globin gene variants in the Pacific, pp. 246-285. In Hill AVS and Serjeantson SW (eds) *The colonization of the Pacific: a genetic trail*. Oxford: Oxford University Press.

Hill AVS, Allsopp CEM, Kwiatkowski D, Anstey NM, Twumasi P, Rowe PA, Bennett S, Brewster D, McMichael AJ and Greenwood BM (1991) Common West African HLA antigens are associated with protection from severe malaria. *Nature (London)* **352**: 595-600.

Horai S and Hayasaka K (1990) Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. *American Journal of Human Genetics* **46**: 828-842.

Horai S, Satta Y, Hayasaka K, Kondo R, Inoue T, Ishida T, Hayashi S and Takahata N (1992) Man's place in hominoidea revealed by mitochondrial DNA genealogy. *Journal of Molecular Evolution* **35**: 32-43.

Huang M-M, Arnheim N and Goodman MF (1992) Extension of base mispairs by *Taq* DNA polymerase: implications for single nucleotide discrimination in PCR. *Nucleic Acids Research* **20**: 4567-4573.

Hubby JL and Lewontin RC (1966) A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics* **54**: 577-594.

Hudson RR (1990) Gene genealogies and the coalescent process, pp. 1-44. In Futuyama D and Antonovics J (eds) *Oxford Surveys in Evolutionary Biology*. Oxford: Oxford University Press.

Hudson RR (1994) How can the low levels of DNA sequence variation in regions of the *Drosophila* genome with low recombination rates be explained? *Proceedings of the National Academy of Sciences USA* **91**: 6815-6818.

Hudson RR, Boos DD and Kaplan NL (1992) A statistical test for detecting geographic subdivision. *Molecular Biology and Evolution* **9**: 138-151.

Hudson RR, Kreitman M and Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**: 153-159.

Hultman T, Ståhl S, Hornes E and Uhlén M (1989) Direct solid phase sequencing of genomic and plasmid DNA using magnetic beads as solid support. *Nucleic Acids Research* **17**: 4937-4946.

Indrak K, Brabec V, Indrakova J, Chrobak L, Sakalova A, Jarosova M, Cermak J, Fei Y-J, Kutlar F, Gu Y-C, Baysal E and Huisman THJ (1992) Molecular characterization of ß-thalassemia in Czechoslovakia. *Human Genetics* **88**: 399-404.

Ingram VM (1957) Gene mutation in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. *Nature (London)* **180**: 326-328.

Innis MA, Myambo KB, Gelfand DH and Brow MAD (1988) DNA sequencing with *Thermus aquaticus* DNA polymerase, and direct sequencing of PCR-amplified DNA. *Proceedings of the National Academy of Sciences USA* **85**: 9436-9440.

Jansen R and Ledley FD (1990a) Disruption of phase during PCR amplification and cloning of heterozygous target sequences. *Nucleic Acids Research* **18**: 5153-5156.

Jansen R and Ledley FD (1990b) Heterozygous mutations at the *mut* locus in fibroblasts with mut^0 methylmalonic acidemia identified by polymerase-chain-reaction cDNA cloning. *American Journal of Human Genetics* **47**: 808-814.

Jeffreys AJ (1979) DNA sequence variants in the $^{G}\gamma$-, $^{A}\gamma$-, δ-, and ß-globin genes of man. *Cell* **18**: 1-10.

Jeffreys AJ, Neumann R and Wilson V (1990) Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying allelic variation and mutation by single molecule analysis. *Cell* **60**: 473-485.

Johnson MJ, Wallace DC, Ferris SD, Rattazzi MC and Cavalli-Sforza LL (1983) Radiation of human mitochondria DNA types analyzed by restriction endonuclease cleavage patterns. *Journal of Molecular Evolution* **19**: 255-271.

Karlsson S and Nienhuis AW (1985) Developmental regulation of human globin genes. *Annual Review of Biochemistry* **54**: 1071-1108.

Kazazian Jr. HH, Chakravarti A, Orkin SH and Antonarakis SE (1983) DNA polymorphisms in the human ß globin gene cluster., pp. 137-146. In Nei M and Koehn RK (eds) *Evolution of genes and proteins*. Sunderland, Massachusetts: Sinauer Associates, Inc.

Kazazian Jr. HH, Orkin SH, Antonarakis SE, Sexton JP, Boehm CD, Goff SC and Waber PG (1984) Molecular characterization of seven ß-thalassemia mutations in Asian Indians. *The EMBO Journal* **3**: 593-596.

Kimura A, Matsunaga E, Takihara Y, Nakamura T, Takagi Y, Lin S-T and Lee H-T (1983) Structural analysis of a ß-thalssemia gene found in Taiwan. *The Journal of Biological Chemistry* **258**: 2748-2749.

Kimura M (1968) Evolutionary rate at the molecular level. *Nature (London)* **217**: 624-626.

Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**: 893-903.

Kimura M (1983) *The neutral theory of molecular evolution*. Cambridge: Cambridge University Press.

Kimura M and Crow JF (1964) The number of alleles that can be maintained in a finite population. *Genetics* **49**: 725-738.

Kitsberg D, Selig S, Keshet I and Cedar H (1993) Replication structure of the human ß-globin gene domain. *Nature (London)* **366**: 588-560.

Klein RG (1992) The archeology of modern human origins. *Evolutionary Anthropology* **1**: 5-14.

Kobayashi K, Jackson MJ, Tick DB, O'Brien WE and Beaudet AL (1990) Heterogeneity of mutations in argninosuccinate synthetase causing human citrullinemia. *The Journal of Biological Chemistry* **265**: 11361-11367.

Kohen G, Philippe N and Godet J (1982) Polymorphism of the Hinf I restriction site located 1 kb 5' to the human §-globin gene. *Human Genetics* **62**: 121-123.

Kreitman M (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature (London)* **304**: 412-417.

Kreitman ME and Aguadé M (1986) Excess polymorphism at the ADH locus in *Drosophila melanogaster*. *Genetics* **114**: 93-110.

Kreitman M and Hudson RR (1991) Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. *Genetics* **127**: 565-582.

Kulozik AE, Wainscoat JS, Serjeant GR, Kar BC, Al-Awamy B, Essan GJF, Falusi AG, Haque SK, Hilali AM, Kate S, Ranasinghe WAEP and Weatherall DJ (1986) Geographical survey of $ß^S$ -globin gene haplotypes: evidence for an independent Asian origin of the sickle-cell mutation. *American Journal of Human Genetics* **39**: 239-244.

Kwiatowski J, Skarecky D, Hernandez S, Pham D, Quijas F and Ayala FJ (1991) High fidelity of the polymerase chain reaction. *Molecular Biology and Evolution* **8**: 884-887.

Kwok S, Kellogg DE, McKinney N, Spasic D, Goda L, Levenson C and Sninsky JJ (1990) Effects of primer-template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies. *Nucleic Acids Research* **18**: 999-1005.

Labie D and Nagel RL (1987) Genetic heterogeneity of sickle mutations. *Acta Haematologica* **78**: 184-185.

Langley CH, MacDonald J, Miyashita N and Aguadé M (1993) Lack of correlation between interspecific divergence and intraspecific polymorphism at the suppressor of forked region in *Drosophila melanogaster* and *Drosophila simulans*. *Proceedings of the National Academy of Sciences USA* **90**: 1800-1803.

Lapouméroulie C, Dunda O, Ducrocq R, Trabuchet G, Mony-Lobé M, Bodo JM, Carnevale P, Labie D, Elion J and Krishnamoorthy R (1992) A novel sickle cell mutation of yet another origin in Africa: the Cameroon type. *Human Genetics* **89**: 333-337.

Lawn RM, Efstratiadis A, O'Connell C and Maniatis T (1980) The nucleotide sequence of the human ß-globin gene. *Cell* **21**: 647-651.

Lee J-S (1991) Alternative dideoxy sequencing of double-stranded DNA by cyclic reactions using *Taq* polymerase. *DNA and Cell Biology* **10**: 67-73.

Lewontin RC (1985) Population genetics. *Annual Review of Genetics* **19**: 81-102.

Lewontin RC and Hubby JL (1966) A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* **54**: 595-609.

Li W-H and Sadler LA (1991) Low nucleotide diversity in man. *Genetics* **129**: 513-523.

Li W-H, Wu C-I and Luo C-C (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *Journal of Molecular Evolution* **21**: 58-71.

Li W-H, Luo C-C and Wu C-I (1985) Evolution of DNA sequences, pp. 1-94. In MacIntyre RJ (ed) *Molecular evolutionary genetics*. New York: Plenum.

Lie-Injo LE, Cai S-P, Wahidijat I, Moeslichan S, Lim ML, Evangelista L, Doherty M and Kan YW (1989) ß-thalassemia mutations in Indonesia and their linkage to ß haplotypes. *American Journal of Human Genetics* **45**: 971-975.

Lincoln SE, Daly MJ and Lander ES (1991) *PRIMER, version 0.5 (computer program)*. Oxford: Molecular Biology VAX.

Livingstone FB (1989) Simulation of the diffusion of the ß-globin variants in the Old World. *Human Biology* **61**: 297-309.

Long JC, Chakravarti A, Boehm CD, Antonarakis S and Kazazian Jr. HH (1990) Phylogeny of human ß-globin haplotypes and its implications for recent human evolution. *American Journal of Physical Anthropology* **81**: 113-130.

Lynch M and Crease TJ (1990) The analysis of population survey data on DNA sequence variation. *Molecular Biology and Evolution* **7**: 377-394.

Maddison DR (1991) African origin of human mitochondrial DNA reexamined. *Systematic Zoology* **40**: 355-363.

Maeda N, Bliska JB and Smithies O (1983) Recombination and balanced chromosome polymorphism suggested by DNA sequences 5' to the human δ-globin gene. *Proceedings of the National Academy of Sciences USA* **80**: 5012-5016.

Marjoram P and Donnelly P (1994) Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. *Genetics* **136**: 673-683.

Marotta CA, Forget BG, Cohen-Solal M, Wilson JT and Weissman SM (1977) Human ß-globin messenger RNA: I. Nucleotide sequences derived from complementary RNA. *Journal of Biological Chemistry* **252**: 5019-5031.

Martin-Campos JM, Cameron JM, Miyashita N and Aguadé M (1992) Intraspecific and interspecific variation at the *y-ac-sc* region of *Drosophila simulans* and *Drosophila melanogaster*. *Genetics* **130**: 805-816.

Martinson JJ, Excoffier L, Boyce AJ and Clegg JB High frequencies of novel α-globin haplotypes in Senegal. *In preparation*.

Matsuno Y, Yamashiro Y, Yamamoto K, Hattori Y, Yamamoto K, Ohba Y and Miyaji T (1992) A possible example of gene conversion with a common ß-thalassemia mutation and Chi sequence present in the ß-globin gene. *Human Genetics* **88**: 357-358.

Maxam AM and Gilbert W (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences USA* **74**: 560-564.

Maxam AM and Gilbert W (1980) Sequencing end-labelled DNA with base-specific chemical cleavages. *Methods in Enzymology* **65**: 499-560.

Maynard Smith J and Haigh J (1974) The hitchhiking effect of a favorable gene. *Genetical Research, Cambridge* **23**: 23-35.

McDonald JH and Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature (London)* **351**: 652-654.

Merriwether DA, Clark AG, Ballinger SW, Schurr TG, Soodyall H, Jenkins T, Sherry ST and Wallace DC (1991) The structure of human mitochondrial DNA variation. *Journal of Molecular Evolution* **33**: 543-555.

Messing J (1983) New M13 vectors for cloning. *Methods in Enzymology* **101**: 20-26.

Miesfeld R, Krystal M and Arnheim N (1981) A member of a new repeated sequence family which is conserved throughout eucaryotic evolution is found between the human δ and ß globin genes. *Nucleic Acids Research* **9**: 5931-5947.

Moschonas N, de Boer E and Flavell RA (1982) The DNA sequence of the 5' flanking region of the human ß-globin gene: evolutionary conservation and polymorphic differences. *Nucleic Acids Research* **10**: 2109-2120.

Mountain JL and Cavalli-Sforza LL (1994) Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. *Proceedings of the National Academy of Sciences USA* **91**: 6515-6519.

Mullis KB, Faloona FA, Scharf SJ, Saiki RK, Horn GT and Erlich HA (1986) Specific enzymatic amplification of DNA *in vitro*: the polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology* **51**: 263-273.

Myers RM, Lumelsky N, Lerman LS and Maniatis T (1985) Detection of single base substitutions in total genomic DNA. *Nature (London)* **313**: 495-498.

Nagamine CM, Chan K and Lau Y-FC (1989) A PCR artifact: generation of heteroduplexes. *American Journal of Human Genetics* **45**: 337-339.

Nei M and Li W-H (1979) Mathematical models for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy Sciences USA* **76**: 5269-5273.

Nei M and Roychoudhury AK (1993) Evolutionary relationships of human populations on a global scale. *Molecular Biology and Evolution* **10**: 927-943.

Newton CR, Graham A, Heptinstall LE, Powell SJ, Summers C, Kalsheker N, Smith JC and Markham AF (1989) Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic Acids Research* **17**: 2503-2516.

Old JM, Heath C, Fitches A, Thein SL, Jefreys AJ, Petrou M, Modell B and Weatherall DJ (1986) Meiotic recombination between two polymorphic restriction sites within the ß globin gene cluster. *Journal of Medical Genetics* **23**: 14-18.

Old JM and Higgs DR (1983) Gene analysis, pp. 74-102. In Weatherall DJ (ed) *The thalassemias: methods in hematology*. London: Churchill Livingstone.

Öner C, Dimovski AJ, Atlay Ç, Gurgey A, Gu YC, Huisman THJ and Lanclos KD (1992) Sequence variations in the 5' hypersensitive site-2 of the locus control region of $ß^S$ chromosomes are associated with different levels of fetal globin in Hemoglobin S homozygotes. *Blood* **79**: 813-819.

Orita M, Iwahana H, Kanazawa H, Hayashi K and Sekiya T (1989a) Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proceedings of the National Academy of Sciences USA* **86**: 2766-2770.

Orita M, Suzuki Y, Sekiya T and Hayashi K (1989b) Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. *Genomics* **5**: 874-879.

Orita M, Sekiya T and Hayashi K (1990) DNA sequence polymorphisms in Alu repeats. *Genomics* **8**: 271-278.

Orkin SH and Kazazian Jr. HH (1984) The mutation and polymorphism of the human ß-globin gene and its surrounding DNA. *Annual Review of Genetics* **18**: 131-171.

Orkin SH, Kazazian Jr. HH, Antonarakis SE, Goff SC, Boehm CD, Sexton JP, Waber PG and Giardina PJV (1982) Linkage of ß-thalassaemia mutations and ß-globin gene polymorphisms with DNA polymorphisms in human ß-globin gene cluster. *Nature (London)* **296**: 627-631.

Pagnier J, Mears JG, Dunda-Belkhodja O, Schaefer-Rego KE, Beldjord C, Nagel RL and Labie D (1984) Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. *Proceedings of the National Academy of Sciences USA* **81**: 1771-1773.

Perutz MF (1990) Frequency of abnormal human haemoglobins caused by C->T transitions in CpG dinucleotides. *Journal of Molecular Biology* **213**: 203-206.

Poncz M, Schwartz E, Ballantine M and Surrey S (1983) Nucleotide sequence analysis of the δß-globin region in humans. *The Journal of Biological Chemistry* **258**: 11599-11609.

Ponticelli AS and Smith GR (1992) Chromosomal context dependence of a eukaryotic recombinational hot spot. *Proceedings of the National Academy of Sciences USA* **89**: 227-231.

Powars DR (1991) $ß^S$-gene-cluster haplotypes in sickle cell anemia: clinical and hematologic features. *Hematology/Oncology Clinics of North America* **5**: 475-493.

Reiss J, Krawczak M, Scholesser M, Wagner M and Cooper DN (1990) The effect of replication errors on the mismatch analysis of PCR-amplified DNA. *Nucleic Acids Research* **18**: 973-978.

Relethford JH and Harpending HC (1994) Craniometric variation, genetic theory, and modern human origins. *American Journal of Physical Anthropology* **95**: 249-270.

Rogers AR and Jorde LB (1994) Genetic evidence on modern human origins. *Human Biology* In press.

Rohlf FJ and Sokal RR (1981) *Statistical tables.* New York: W. H. Freeman and Company.

Rossiter BJF and Caskey CT (1990) Molecular scanning methods of mutation detection. *The Journal of Biological Chemistry* **265**: 12753-12756.

Ruano G, Kidd KK and Stephens JC (1990) Haplotype of multiple polymorphisms resolved by enzymatic amplification of single DNA molecules. *Proceedings of the National Academy of Sciences USA* **87**: 6296-6300.

Ruvolo M, Zehr S, von Dornum M, Pan D, Chang B and Lin J (1993) Mitochondrial COII sequences and modern human origins. *Molecular Biology and Evolution* **10**: 1115-1135.

Saiki RK, Scharf SJ, Faloona FA, Mullis KB, Horn GT, Erlich HA and Arnheim N (1985) Enzymatic amplification of beta-globin sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**: 1350-1354.

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB and Erlich HA (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**: 487-491.

Saitou N and Omoto K (1987) Time and place of human origins from mtDNA data. *Nature (London)* **327**: 288.

Saitou N and Ueda S (1994) Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. *Molecular Biology and Evolution* **11**: 504-512.

Sambrook J, Fritsch EF and Maniatis T (1989) *Molecular cloning: a laboratory manual.* Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

Sanger F, Nicklen S and Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences USA* **74**: 5463-5467.

Savatier P, Trabuchet G, Faure C, Chebloune Y, Gouy M, Verdier G and Nigon VM (1985) Evolution of the primate ß-globin gene region: high rate of variation in CpG dinucleotides and in short repeated sequences between man and chimpanzee. *Journal of Molecular Biology* **182**: 21-29.

Schaeffer SW and Miller EL (1991) Nucleotide sequence analysis of *Adh* genes estimates the time of geographic isolation of the Bogota population of *Drosophila pseudoobscura*. *Proceedings of the National Academy of Sciences USA* **88**: 6097-6101.

Schaeffer SW and Miller EL (1992) Molecular population genetics of an electrophoretically monomorphic protein in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. *Genetics* **132**: 163-178.

Scharf SJ, Long CM and Erlich HA (1988) Sequence analysis of the HLA-DR beta and HLA-DQ beta loci from three *Pemphigus vulgaris* patients. *Human Immunology* **22**: 61-69.

Schultes NP and Szostak JW (1991) A poly(dA-dT) tract is a component of the recombination initiation site at the *ARG4* locus in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology* **11**: 322-328.

Semenza GL, Delgrosso K, Poncz M, Malladi P, Schwartz E and Surrey S (1984a) The silent carrier allele: ß thalassemia without a mutation in the ß-globin gene or its immediate flanking regions. *Cell* **39**: 123-128.

Semenza GL, Malladi P, Surrey S, Delgrosso K, Poncz M and Schwartz E (1984b) Detection of a novel DNA polymorphism in the ß-globin gene cluster. *The Journal of Biological Chemistry* **259**: 6045-6048.

Shenkar R, Shen M and Arnheim N (1991) DNase I-hypersensitive sites and transcription factor-binding motifs within the mouse Eß meiotic recombination hot spot. *Molecular and Cellular Biology* **11**: 1813-1819.

Shuldiner AR, Scott LA and Roth J (1990) PCR-induced (ligase-free) subcloning: a rapid reliable method to subclone polymerase chain reaction (PCR) products. *Nucleic Acids Research* **18**: 1920.

Slatkin M (1985) Rare alleles as indicators of gene flow. *Evolution* **39**: 53-65.

Smith GP (1994) The progeny of sexual PCR. *Nature (London)* **370**: 324-325.

Spence SE, Pergolizzi RG, Donovan-Peluso M, Kosche KA, Dobkin CS and Bank A (1982) Five nucleotide changes in the large intervening sequence of a ß globin gene in a $ß^+$ thalassemia patient. *Nucleic Acids Research* **10**: 1283-1294.

Stamatoyannopoulos G and Nute PE (1982) De novo mutations producing unstable Hbs or Hbs M: II. Direct estimates of minimum nucleotide mutation rates in man. *Human Genetics* **60**: 181-188.

Stemmer WPC (1994) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature (London)* **370**: 389-391.

Stephens JC and Nei M (1985) Phylogenetic analysis of polymorphic DNA sequences at the Adh locus in *Drosophila melanogaster* and its sibling species. *Journal of Molecular Evolution* **22**: 289-300.

Stewart C-B (1993) The powers and pitfalls of parsimony. *Nature (London)* **361**: 603-607.

Stoflet ES, Koeberl DD, Sarkar G and Sommer SS (1988) Genomic amplification with transcript sequencing. *Science* **239**: 491-494.

Stoneking M, Jorde LB, Bhatia K and Wilson AC (1990) Geographic variation in human mitochondrial DNA from Papua New Guinea. *Genetics* **124**: 717-733.

Stringer CB (1992) Reconstructing recent human evolution. *Philosophical Transactions of the Royal Society, B* **337**: 217-224.

Stringer CB and Andrews P (1988) Genetic and fossil evidence for the origin of modern humans. *Science* **239**: 1263-1268.

Stringer C and Bräuer G (1994) Methods, misreading, and bias. *American Anthropologist* **96**: 416-424.

Strobeck C (1987) Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics* **117**: 149-155.

Strobeck C and Morgan K (1978) The effect of intragenic recombination on the number of alleles in a finite population. *Genetics* **88**: 829-844.

Sutton M, Bouhassira EE and Nagel RL (1989) Polymerase chain reaction amplification applied to the determination of ß-like globin gene cluster haplotypes. *American Journal of Hematology* **32**: 66-69.

Swisher CC, Curtis GH, Jacob T, Getty AG, Suprijo A and Widiasmoro (1994) Age of the earliest known hominids in Java, Indonesia. *Science* **263**: 1118-1121.

Swofford DL (1991) *PAUP: phylogenetic analysis using parsimony, version 3.1 (computer program)*. Champaign, Illinois: Illinois Natural History Survey.

Symington LS, Brown A, Oliver SG, Greenwell P and Petes TD (1991) Genetic analysis of a meiotic recombination hotspot on chromosome *III* of *Saccharomyces cerevisiae*. *Genetics* **128**: 717-727.

Szostak JW, Orr-Weaver TL, Rothstein RJ and Stahl FW (1983) The double-strand-break repair model for recombination. *Cell* **33**: 25-35.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**: 437-460.

Tajima F (1989a) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585-595.

Tajima F (1989b) The effect of change in population size on DNA polymorphism. *Genetics* **123**: 597-601.

Takahata N (1993) Allelic genealogy and human evolution. *Molecular Biology and Evolution* **10**: 2-22.

Templeton AR (1992) Human origins and analysis of mitochondrial DNA sequences. *Science* **255**: 737.

Templeton AR (1993) The "Eve" hypothesis: a genetic critique and reanalysis. *American Anthropologist* **95**: 51-72.

Thein SL (1993) ß-thalassaemia. *Bailliére's Clinical Haematology* **6**: 151-175.

Thein SL and Hinton J (1991) A simple and rapid method of direct sequencing using Dynabeads. *British Journal of Haematology* **79**: 113-115.

Thorne AG and Wolpoff MH (1992) The multiregional evolution of humans. *Scientific American* **266**: 28-33.

Trabuchet G, Elion J, Baudot G, Pagnier J, Bouhass R, Nigon VM, Labie D and Krishnamoorthy R (1991) Origin and spread of ß-globin gene mutations in India, Africa, and Mediterranea: analysis of the 5' flanking and intragenic sequences of $ß^S$ and $ß^C$ genes. *Human Biology* **63**: 241-252.

Treco D and Arnheim N (1986) The evolutionarily conserved repetitive sequence d(TG-AC)$_n$ promotes reciprocal exchange and generates unusual recombinant tetrads during yeast meiosis. *Molecular and Cellular Biology* **6**: 3934-3947.

Treco D, Thomas B and Arnheim N (1985) Recombination hot spot in the human ß-globin gene cluster: meiotic recombination of human DNA fragments in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology* **5**: 2029-2038.

Treisman R, Orkin SH and Maniatis T (1983) Specific transcription and RNA splicing defects in five cloned ß-thalassaemia genes. *Nature (London)* **302**: 591-596.

van Delden W (1982) The alcohol dehydrogenase polymorphism in *Drosophila melanogaster*. Selection at an enzyme locus. *Evolutionary Biology* **15**: 187-222.

Vigilant L, Pennington R, Harpending H, Kocher TD and Wilson AC (1989) Mitochondrial DNA sequences in single hairs from a southern African population. *Proceedings of the National Academy of Sciences USA* **86**: 9350-9354.

Vigilant L, Stoneking M, Harpending H, Hawkes K and Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. *Science* **253**: 1503-1507.

Waddle DM (1994) Matrix correlation tests support a single origin for modern humans. *Nature (London)* **368**: 452-454.

Wahls WP, Wallace LJ and Moore PD (1990) The Z-DNA motif $d(TG)_{30}$ promotes reception of information during gene conversion events while stimulating homologous recombination in human cells in culture. *Molecular and Cellular Biology* **10**: 785-793.

Wainscoat JS, Hill AVS, Boyce AJ, Flint J, Hernandez M, Thein SL, Old JM, Lynch JR, Falusi AG, Weatherall DJ and Clegg JB (1986) Evolutionary relationships of human populations from an analysis of nuclear DNA polymorphisms. *Nature (London)* **319**: 491-493.

Ward MA, Skandalis A, Glickman BW and Grosovsky AJ (1989) Rapid generation of specific single-stranded template DNA from PCR-amplified material. *Nucleic Acids Research* **17**: 8394.

Watt WB (1972) Intragenic recombination as a source of population genetic variability. *American Naturalist* **106**: 737-753.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**: 256-276.

White MB, Carvalho M, Derse D, O'Brien SJ and Dean M (1992) Detecting single base substitutions as heteroduplex polymorphisms. *Genomics* **12**: 301-306.

Wijsman EM and Cavalli-Sforza LL (1984) Migration and genetic population structure with special reference to humans. *Annual Review of Ecology and Systematics* **15**: 279-301.

Wolpoff M (1989) Multiregional evolution: the fossil alternative to Eden, pp. 62-108. In Mellars P and Stringer C (eds) *The human revolution: behavioral and biological perspectives on the origins of modern humans*. Princeton, NJ: Princeton University Press.

Wong C, Dowling CE, Saiki RK, Higuchi RG, Erlich HA and Kazazian Jr. HH (1987) Characterization of ß-thalassemia mutations using direct genomic sequencing of amplified single copy DNA. *Nature (London)* **330**: 384-386.

Wood N, Tyfield L and Bidwell J (1993) Rapid classification of phenylketonuria genotypes by analysis of heteroduplexes generated by PCR-amplifiable synthetic DNA. *Human Mutation* **2**: 131-137.

Wright S (1931) Evolution in Mendelian populations. *Genetics* **16**: 97-159.

Wright S (1951) The genetical structure of populations. *Annals of Eugenics* **15**: 323-354.

Wrischnik LA, Higuchi RG, Stoneking M, Erlich HA, Arnheim N and Wilson AC (1987) Length mutations in human mitochondrial DNA: direct sequencing of enzymatically amplified DNA. *Nucleic Acids Research* **15**: 529-542.

Wu DY, Ugozzoli L, Pal BK and Wallace RB (1989) Allele-specific enzymatic amplication of ß-globin genomic DNA for diagnosis of sickle cell anemia. *Proceedings of the National Academy of Sciences USA* **86**: 2757-2760.

Würm SA (1983) Linguistic prehistory in the New Guinea area. *Journal of Human Evolution* **12**: 25-35.

Zhang J-Z, Cai S-P, He X, Lin H-X, Lin H-J, Huang Z-G, Chehab FF and Kan YW (1988) Molecular basis of ß thalassemia in South China. *Human Genetics* **78**: 37-40.